

FACILITATING MECHANISTIC STUDIES OF PRE-MRNA SPLICING VIA A  
NOVEL TARGETED SEQUENCING APPROACH

A Dissertation

Presented to the Faculty of the Graduate School  
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy

by

Hansen Xu

December 2017

© 2017 Hansen Xu

# FACILITATING MECHANISTIC STUDIES OF PRE-MRNA SPLICING VIA A NOVEL TARGETED SEQUENCING APPROACH

Hansen, Ph. D.

Cornell University 2017

Most eukaryotic genes have their protein coding sequences interrupted by non-coding introns, which must be removed from nascent pre-mRNAs by the spliceosome to generate a translatable mRNA. The past few decades have been marked by a significant increase in our appreciation for the central role that splicing plays in regulating eukaryotic gene expression. Nevertheless, the mechanisms by which this process is normally regulated, and can be mis-regulated with pathological consequences, remain poorly understood.

Next-generation sequencing (NGS) technologies have had a profound effect on our understanding of pre-mRNA splicing. Yet in spite of the power presented by this approach, it is less widely appreciated that the depth of sequencing necessary to quantitatively detect many splicing isoforms is significantly higher than most RNA-Seq experiments generate. Here I present the development and implementation of a novel sequencing method designed to harness the quantitative power of sequencing while focusing it on user-selected splice junctions of interest through targeting for introns at the reverse transcription step. This approach can dramatically enrich the fraction of reads in each sequencing experiment that are informative about splicing status, and in doing so enable a significant increase in the precision with which changes in splicing can be detected, all while decreasing experimental costs.

## BIOGRAPHICAL SKETCH

Hansen Xu was born in China and immigrated with his parents to Canada at the age of 8. Growing up, he had always known that he had an interest in the sciences, and it led to his pursuit of an undergraduate degree in chemistry and biochemistry at Cornell University. During his undergraduate career, Hansen worked in the Department of Chemical Engineering under the guidance of Dr. Matthew DeLisa and worked on an array of topics ranging from scFv optimization to metabolic engineering in *E. coli*.

For reasons that remain unclear even to himself, he chose to stay in Ithaca to obtain his Ph.D. at Cornell University, working under Dr. Jeffrey Pleiss studying the mechanisms regulating splicing in *S. cerevisiae*. Even though many scientific obstacles were encountered along the way, after 11 long and cold winters he is finally ready to graduate and looks forward to reuniting with his family and starting his new life in Shanghai, China.

This work is dedicated to my loving parents, family members and all my friends  
who have supported me over the years,  
and without whom I'd probably have graduated 2 years earlier.

## ACKNOWLEDGMENTS

I want to thank my advisor, Dr. Jeffrey Pleiss, for his guidance through my training as a scientist. Jeff has always offered advice on anything you'd have asked of him, both scientific and life related.

I want to thank the members of my committee Dr. Eric Alani and Dr. Ailong Ke, both of whom made sure I was on the right track with my research plans and offered valuable second opinions.

I'd like to thank Dr. Andrew Grimson and members of his lab for all the valuable discussions we've had as we've had shared group meetings for the past 6 years.

I would like to thank all past and current members of Pleiss lab. Especially Ben and Zach, whom are co-authors with me on this paper, for providing valuable help in getting the computational part of the experiments worked through. Madhura, with whom I've shared the Mud1 project upon joining the Pleiss lab. Laura, for her guidance in mentoring me when I first joined the lab and her contributions to the Bdf1 project. Mike for always providing fresh organic eggs when I asked him.

And finally, I'd like to thank my family. Mom and Dad, thank you for all your love and support over the past 6 years. Couldn't have done this without you guys!

## TABLE OF CONTENTS

<b>Preface .....</b>	<b>1</b>
Abstract .....	3
Biographical sketch.....	4
Dedications .....	5
Acknowledgements.....	6
Table of contents.....	7
<b>1. Introduction .....</b>	<b>9</b>
1.1 Introduction to splicing .....	9
1.2 The spliceosome is the core splicing machinery.....	10
1.3 Structure of the spliceosome.....	18
1.4 Splicing is an important and essential process in biology .....	23
1.5 Splicing is coupled to transcription.....	23
1.6 Yeast as a model organism to study splicing .....	24
1.7 Historical methods used to monitor splicing .....	25
1.8 Current methods used to monitor splicing.....	26
1.9 Motivation for this dissertation.....	31
<b>2. Splice-Seq as a novel targeted sequencing approach to study pre-mRNA splicing.....</b>	<b>32</b>
2.1 Abstract .....	32
2.2 Introduction.....	34
2.3 Results.....	36
2.4 Discussion .....	59
2.5 Materials and methods .....	61
<b>3. Deciphering a role for the chromatin remodeling factor Bdf1 in pre-mRNA splicing.....</b>	<b>67</b>
3.1 Introduction.....	67
3.2 Materials and methods .....	75
3.3 Results and discussion .....	78
3.3.1 ChIP-Seq of Bdf1 mutants.....	78
3.3.2 Forward genetics screen of Bdf1 mutants .....	81
<b>Future Directions.....</b>	<b>87</b>

<b>Credits .....</b>	<b>91</b>
<b>Bibliography.....</b>	<b>92</b>



## Chapter 1: Introduction

### 1.1: Introduction to splicing

All life rely on genetic information found in genes to properly function. The central dogma of biology, which explains the transfer of genetic information, states that in order to carry out a gene's designated cellular role, genetic information flows from DNA to RNA in a process called transcription and subsequently to proteins in a processed called translation (Yung and Primm 2015, Figure 1.1).

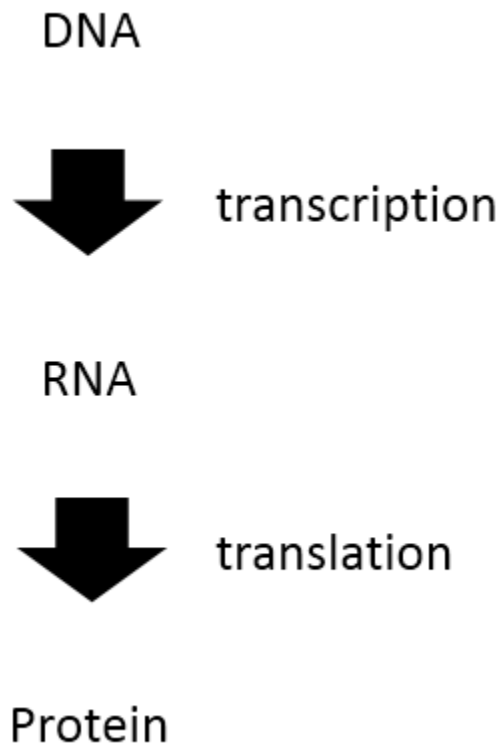


Figure 1.1: The central dogma of biology. Arrows represent flow of genetic information

Genes are present in the genome in the form of DNA nucleotides that are used as a template to transcribe RNA transcripts. Subsequently the RNA can act as a

translation template to create the protein product for that gene. This flow of genetic information is collectively coined gene expression and must be carefully regulated by the organism to ensure the proper amount of protein products are synthesized (Rockman and Kruglyak 2006). There are many ways in which gene expression can be regulated, and one such example is through a process called splicing (Long and Caceres 2009). Many eukaryotic organisms contain non-coding intervening sequences (introns) in their genes in addition to coding sequences (exons). In general, intervening sequences do not contain useful information for protein synthesis in translation and must be removed for the pre-messenger RNAs (pre-mRNAs) to be altered into a mature messenger RNA (mRNA), the substrate for translation (William Roy and Gilbert 2006). This process of intron removal and exon joining is called splicing, an important mechanism for regulating gene expression (Hughes 2006).

### **1.2: The spliceosome is the core splicing machinery**

Splicing at its core is the combination of two sequential transesterification reactions that first removes the intron and subsequently joins the exons (Chanfreau and Jacquier 1996). As an example, the simplest form of splicing, in which two exons are separated by a single intron is illustrated in (Figure 1.2). The 5' and 3' end boundaries of the intron are termed the 5' splice site (5' SS) and the 3' splice site (3' SS), respectively.

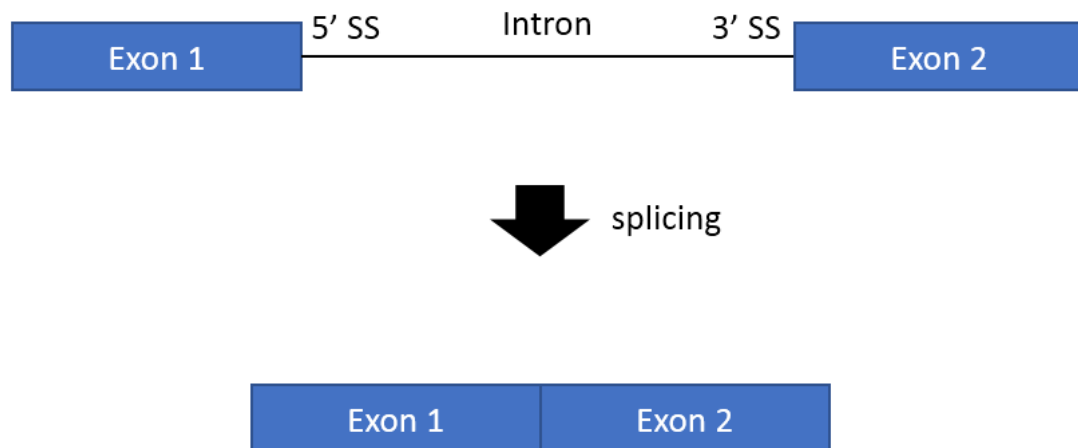


Figure 1.2: Simple representation of splicing. Blue boxes represent exons, black line represents the intron. Locations of the 5' and 3' splice sites are labeled.

Splicing is often more complex than the simple removal of an intron. For example, a gene can retain the intron in the mRNA instead of splicing it out in a process called intron retention (Jung et al. 2015). Alternatively, the intron can be removed, but with a different 5' SS and/or 3' SS than the canonical one. Additionally, in genes where 2 or more introns exist, different exons can be skipped when splicing occurs, resulting in very different mRNAs from the canonically spliced transcript (Figure 1.3). These different forms of splicing are collectively called alternative splicing (Chen and Manley 2009) and add further control in regulating gene expression by controlling the protein product formed from a gene.

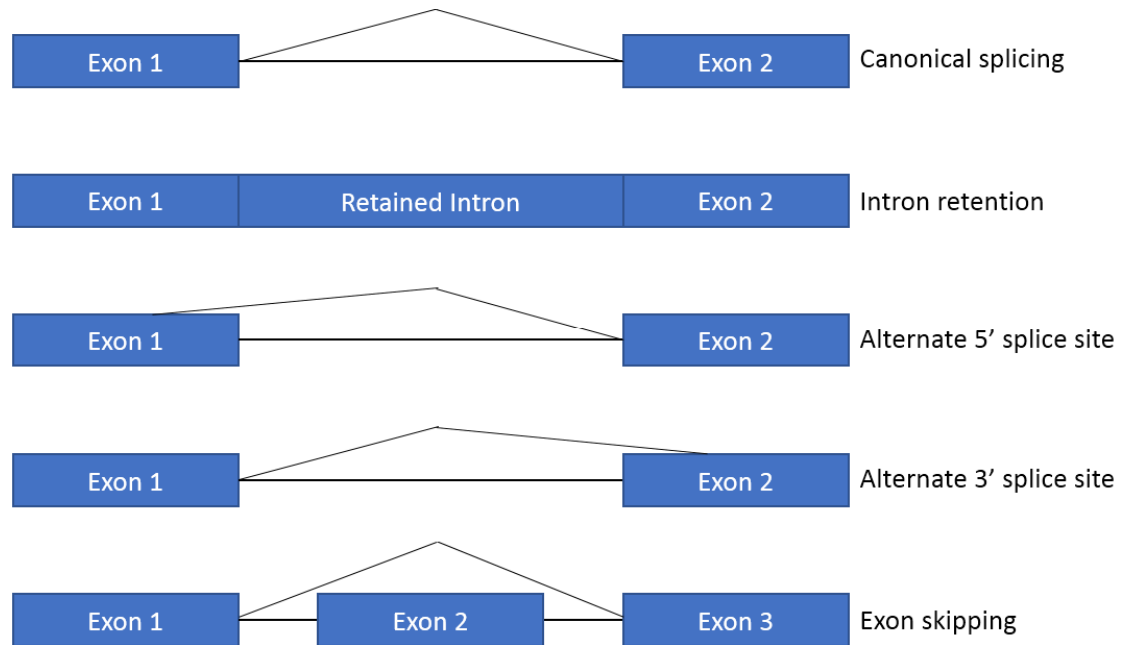


Figure 1.3: Examples of alternative splicing. Lines above transcript signifies positions where exons are joined.

Although the underlying transesterifications behind splicing are rather simple chemical reactions on their own, to ensure high fidelity and efficiency, the molecular machinery required in this process, the spliceosome, is massive and complex (Matera and Wang 2014). The conserved eukaryotic spliceosomal machinery is composed of five small nuclear ribonucleoproteins (snRNPs) named U1, U2, U4, U5 and U6. Each of the snRNPs contain a unique small nuclear RNA (snRNA) along with a set of Sm or Lsm proteins and a varying number of other particle specific proteins (Will and Luhrmann 2011). The massive complex does not have pre-formed active sites, but rather must be assembled in a stepwise fashion onto each target transcript (Chiou and Lynch 2014). Assembly begins with binding of U1 snRNP at the 5' splice site (5' SS) of the pre-mRNA. The U2 snRNP subsequently binds to the branch point (BP) of the pre-mRNA, forming the A complex, followed by the addition of the pre-assembled U4/6-U5 tri-snRNP, forming the pre-catalytic B complex. Upon the U4/6-U5 tri-snRNP

binding, major RNA-RNA and RNA-protein rearrangements occur and U1 and U4 snRNPs are destabilized and the spliceosome is activated for catalysis, giving rise to the  $B^{\text{act}}$  complex. (Staley and Guthrie 1998). Catalytic activation by the DEAH-box RNA helicase Prp2 generates the  $B^*$  complex which catalyzes the first of the two steps of splicing, resulting in the formation of the C complex. The C complex then completes the second step of splicing, resulting in the formation of an excised intron lariat while joining the exons in question. Spliceosomal components are then recycled to be used on another intron (Will and Luhrmann 2011, Figure 1.4).

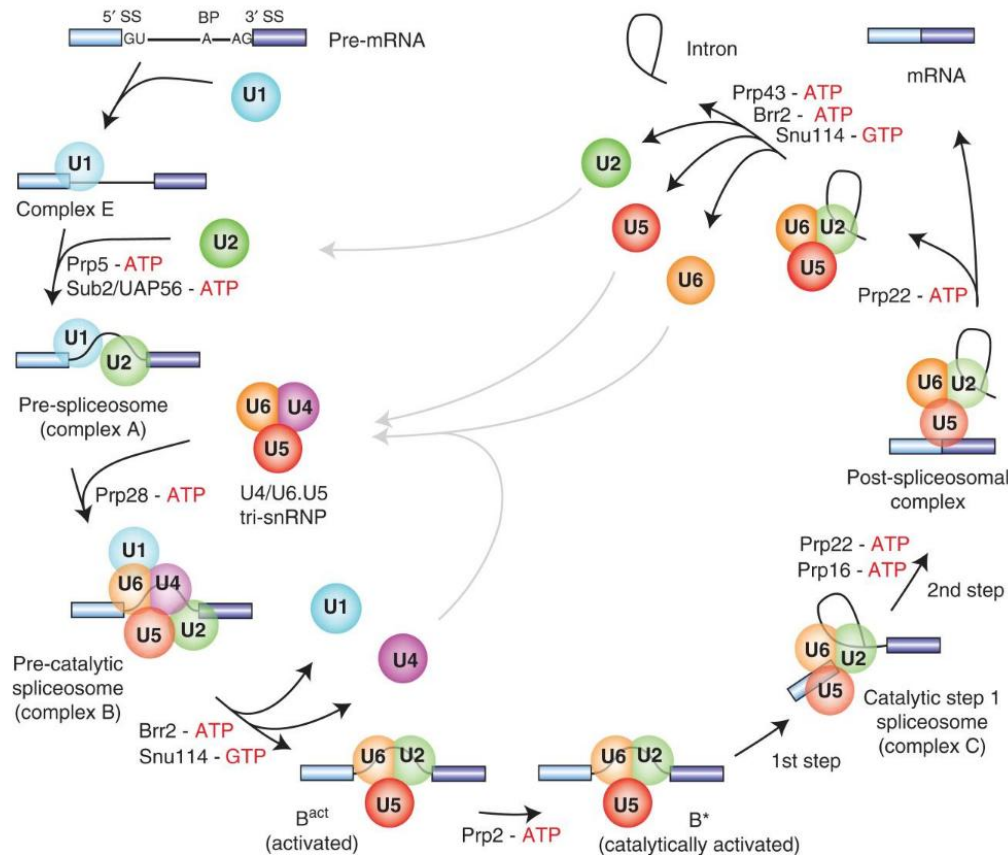


Figure 1.4: Illustration of the detailed process of spliceosome assembly and steps of splicing. Circles represent the snRNPs. Adapted from Will and Luhrmann 2011.

At first glance, it is apparent that U1 and U2 snRNPs find their respective targets,

the 5' SS and BP respectively, through complementary base pairing between their snRNAs and the sequences of the target sites. However, the complementarity alone does not explain the high observed fidelity of splicing (Kim et al. 2017). Some mutations in the 5'SS can be suppressed by matching mutation in the corresponding snRNA, but other mutations cannot, suggesting base pairing between U1 and the 5'SS is necessary but not sufficient for the splicing of mRNA precursors (Zhuang and Weiner 1986). The consensus sequence degeneracy also varies greatly between organisms. For example, even though the consensus sequence at the 5'SS is quite conserved in yeast, it is very degenerate in humans, where only a GT (GU in RNA) is absolutely required (Figure 1.5, Carmel 2004). This degeneracy brings up interesting questions of how fidelity is maintained and cryptic 5'SSs are avoided, as many non-splice site GUs appear through genes by random chance. The currently accepted hypothesis in the field is that the degeneracy allows for additional opportunities for extra factors to regulate splicing, an ongoing topic of research in the field (Crotti and Horowitz 2009).

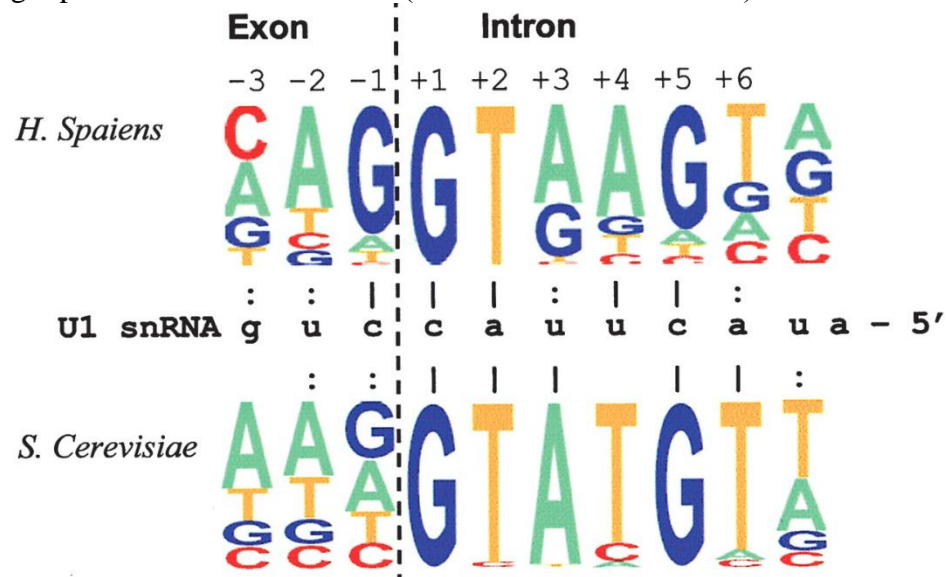


Figure 1.5: Human 5' splice sites are more degenerate than budding yeast.

“WebLogo” (Crooks et al. 2004) representation of 5' splice sites for humans and budding yeast shown. U1 snRNA sequence shown in middle. Letter height represents

frequency of basepair occurrences at the indicated location. Dotted line represent the boundary between exon and intron.

A simplification of the complex splicing process can be reduced to just the two core chemical steps, both of which are trans-esterifications. To illustrate this, consider a simple single intron containing transcript. The first chemical step happens when the 2' hydroxyl group of the conserved adenosine residue at the branchpoint attacks the conserved guanine of the 5' splice site (Maschhoff and Padgett 1993). This results in the formation of an unusual 2' - 5' phosphodiester RNA lariat structure between the branchpoint adenosine and the 5' splice site guanine, and cleaves the 5' exon1-intron junction, leaving exon 1 with a free 3' hydroxyl, and exon 2 connected to the intron in what is termed a lariat intermediate. The second chemical step (Horowitz 2012) happens when the 3' hydroxyl of exon1 attacks the scissile phosphodiester bond of the conserved guanine at the 3' splice site at the end of the intron (Podar, Perlman, and Padgett 1998). This results in the joining of exon 1 and exon 2, and the release of the lariat intron structure (Figure 1.6). It is important in studying splicing that the intermediate products can be monitored to evaluate its efficiency and fidelity (Coombes and Boeke 2005).

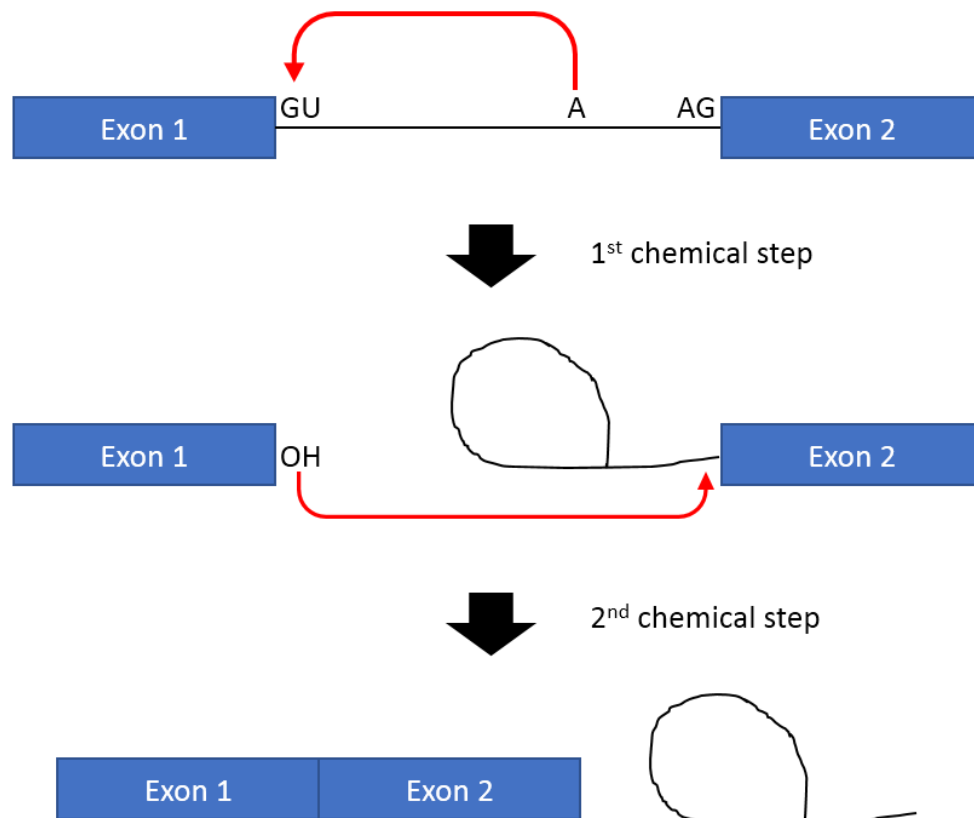


Figure 1.6: Illustration of both chemical steps of splicing. Red arrow indicates the nucleophilic attack of the transesterification reactions.

Over the past 20 years or so, many of the factors important for the two chemical steps of splicing have been identified and characterized to detail (Chen and Cheng 2012, Änkö 2014, Anczuków and Krainer 2016). Importantly, mutations in many of these factors have also been identified and well characterized and are known to cause perturbations in the splicing pathway (Vijayraghavan, Company, and Abelson 1989, Dreumont and Séraphin 2013). Particularly, these mutants can be taken advantage of to study the chemical steps of splicing in great detail. In this dissertation, mutants of two important such factors, Prp2 and Prp16, which function prior to the first and second



chemical steps of splicing, respectively (Figure 1.4, Will and Luhrmann 2011), have been leveraged to examine different aspects of the splicing pathway.

Prp2 is an RNA-dependent DExD/H-box ATPase that is required for the activation of the spliceosome before the first transesterification step in RNA splicing (Kim and Lin 1996). Prp2 binds to a precatalytic spliceosome prior to the first step of splicing, and in concert with ATP hydrolysis causes a change in the overall spliceosome structure, “activating” the spliceosome for the first chemical step of splicing (Kim and Lin 1996, Liu and Cheng 2012). Prp2 is also important for destabilizing RNA elements that comprise the catalytic core of the spliceosome, both to proofread spliceosome activation and to promote reconfiguration of the spliceosome to a fully competent, catalytic formation (Wlodaver and Staley 2014). Mutants that were identified in the Prp2 protein were found to accumulate unspliced RNAs from the vast majority of intron-containing genes (Gencheva et al. 2010). This is not surprising as one would imagine that blocking an important protein for the first chemical step of splicing would block the first step of splicing, and cause an accumulation of the reactants for that chemical reaction, the fully unspliced pre-mRNAs. In this dissertation, we make use of a *prp2-1* temperature sensitive mutant in order to assess the effect of loss of Prp2 protein function on splicing (Kim et al. 1999). While strains harboring the *prp2-1* mutation are able to grow with nearly wild type efficiency at 25 °C, they are unable to support growth at 37 °C (Kim et al. 1999). We have previously shown that genome-wide defects in pre-mRNA splicing are apparent even after short times exposed to elevated temperature (Pleiss et al. 2007), making this mutant an excellent one to use to evaluate the consequences of blocking the first chemical step of splicing.

Prp16, on the other hand, is a DEAH-box RNA helicase involved in the second catalytic step of splicing and in exon ligation (Umen and Guthrie 1995, Ohrt et al. 2013). The protein has been shown to also exhibit RNA-dependent ATPase activity which is

required for the second catalytic step of splicing (Schwer and Guthrie 1991). Furthermore, hPrp16 was shown via mass-spectroscopy to be highly associated with the activated spliceosomal C complex only, suggesting that it plays a role in the second step of splicing but is neither present nor necessary for the first (Schmidt et al., 2014). *In vitro* studies have shown Prp16 to promote ATP-dependent RNA unwinding activity, in a sequence-independent fashion, suggesting the mechanism of function of Prp16 lies in the disruption of a duplexed RNA structure in the spliceosome (Wang, Wagner, and Guthrie 1998). Mutants that were identified in the Prp16 protein were found to accumulate several lariat intermediates (Vijayraghavan, Company, and Abelson 1989). This is in line with the expectations that having a block in the second chemical step of splicing would cause an accumulation of a first step of splicing product accumulation. In this dissertation, we make use of a *prp16-302* temperature sensitive mutant in order to assess the effect of loss of Prp16 protein function on splicing (Villa and Guthrie 2005). Yeast strains carrying the cold-sensitive allele *prp16-302* stall the release of the Prp16 protein at low temperatures and becomes unable to support growth at the non-permissive temperature of 16 °C. At the permissive temperature of 30 °C, growth appears normal but it is unclear whether any silent molecular defects are present but hidden (unpublished data). At the non-permissive temperature, an increase in the lariat intermediates can be observed in conjunction with a decrease in the amount of mature mRNA (Villa and Guthrie 2005), further suggesting Prp16's role in the second step of splicing, and making this mutant an excellent one to study effects from a defective second step of splicing.

### **1.3: Structure of the spliceosome**

As both the conformation and composition of the spliceosome are highly dynamic, understanding the 3-dimensional structure of the spliceosome at high

resolution has been a major challenge for the field (Will and Luhrmann 2011). Electron microscopy (Deckert et al. 2006, Fabrizio et al. 2009, Herold et al. 2009), NMR (Spadaccini et al. 2006) and x-ray crystallography (Oubridge et al. 1994, Price, Evans, and Nagai 1998, Vidovic et al. 2000, Kambach et al. 1999, Ritchie, Schellenberg, and MacMillan 2009) techniques have been employed to gain structural insights into many splicing factors either alone or coupled to their binding partners, but until recently, the structure of the whole and active spliceosome remain elusive. During 2015 - 2016, a series of papers came out that identified the structure of the whole spliceosome instead of single components.

One of the first spliceosomal subcomplexes for which a near atomic-level structure was solved was the U4/U6.U5 tri-snRNP of *S. cerevisiae*. At ~1.5 metadaltons, the tri-snRNP was structurally determined via single-particle cryo-electron microscopy (Cryo-EM) at an average resolution of 5.9 angstroms (Nguyen et al., 2015). Though incredibly insightful into the activation process and the active site of the spliceosome at the time, the resolution was not high enough to resolve many structural details. Shortly after, a high resolution structure was determined of an *S. pombe* spliceosome at an average resolution of 3.6 angstroms by Cryo-EM. This structure contained the U2 and U5 snRNPs, the Prp nineteen complex (NTC), the NTC-related (NTR) complex, U6 snRNA and an RNA intron lariat. This structure had a combined molecular mass of ~1.3 megadaltons, and represented a mixture of the previously described B<sup>act</sup>, B\*, and C complexes, as well as a post-catalytic (P) and intron lariat spliceosomal (ILS) complexes (Yan et al., 2015). This spliceosomal structure represented the first atomic resolution model of an intact, functional spliceosome and prompted a slew of other, more purified spliceosomal structures to be solved.

One such spliceosome structure subsequently solved was the *S. cerevisiae* tri-snRNP complex U4/U6.U5. By use of Cryo-EM, an overall resolution of 3.8 angstroms

was achieved (Figure 1.7, Wan et al. 2016). This massive, ~1 megadalton structure offered invaluable insights into the activation process of the spliceosomal ribozyme, confirming many known roles of splicing factors in addition to discovery of a few unknown factors. For example, the Dib1 protein in *S. cerevisiae* is postulated to play an important role in pre-mRNA splicing as suggested by both its central location in the tri-snRNP and its association with U4 snRNA, the pre-mRNA, Prp31 and the central protein Prp8, even though its function remains to be determined.

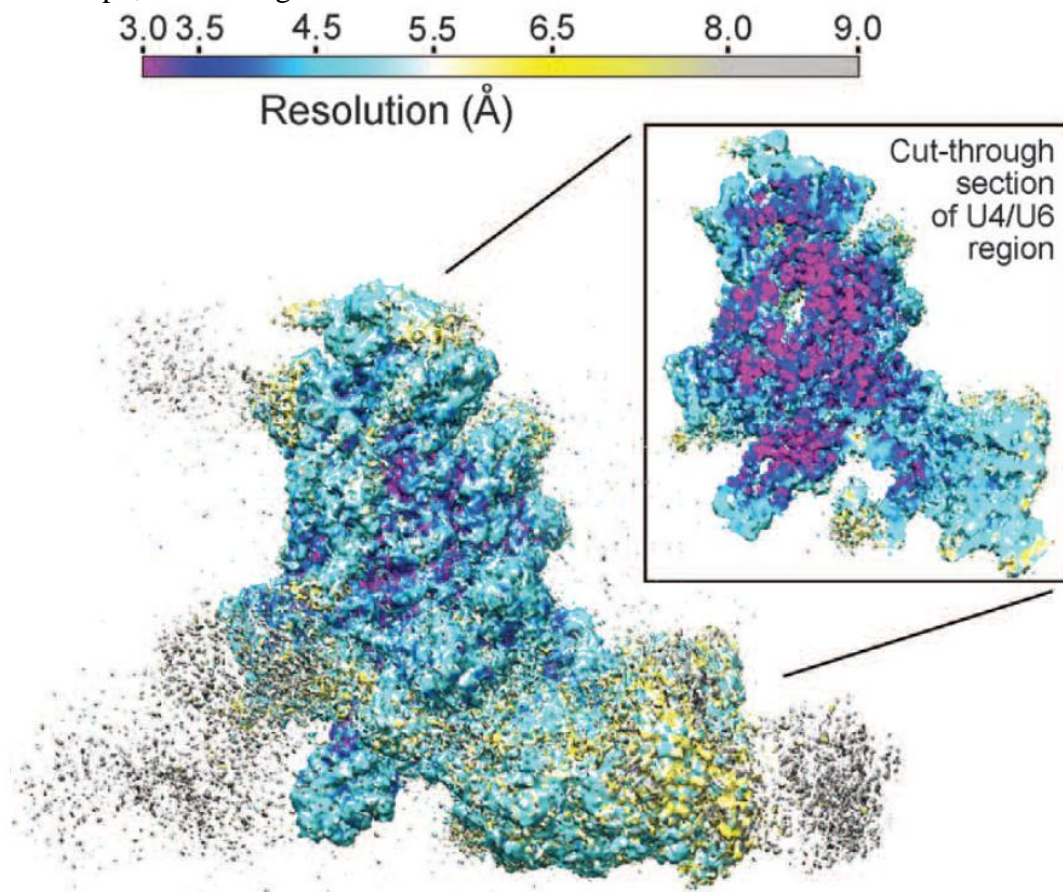


Figure 1.7: cryo-EM map for the yeast U4/U6.U5 tri-snRNP. Colors indicate resolution. Adapted from Wan et al. 2016.

Several months later, the atomic structure of the *S. cerevisiae* activated spliceosome obtained by Cryo-EM, also known as the B<sup>act</sup> complex, had also been

reported, with an average resolution of 3.52 angstroms (Figure 1.8, Yan et al. 2016). The structure, including the 3 snRNAs and the pre-mRNA, has a combined molecular mass of ~1.6 megadaltons. The structural features identified from this structure confirmed the hypothesis in the field that the active site in the B<sup>act</sup> complex has yet to be converted to a catalytically active site. There is a catalytic Mg<sup>2+</sup> atom that is yet to be loaded into the active site location, which would require a minor conformation change in the U6 snRNA to happen. Importantly, the central component of the spliceosome, Prp8, which binds to all four RNA elements, displayed excellent resolution for the structure. This led to the discovery of a switch loop region (residues 1402 to 1439) in the Prp8 protein which undergoes 180 ° flips between its conformation in the U4/U6.U5 tri-snRNP and that of the B<sup>act</sup> conformation (Yan et al. 2016). This rather important discovery led to the hypothesis that the switch loop plays an important function: that its close interaction with the splicing factor Cwc21 preserves the connection between Cwc21 and the 5' exonic sequences through both steps of splicing. This offered valuable insights into the way which spliceosomes are assembled and activated for catalysis.

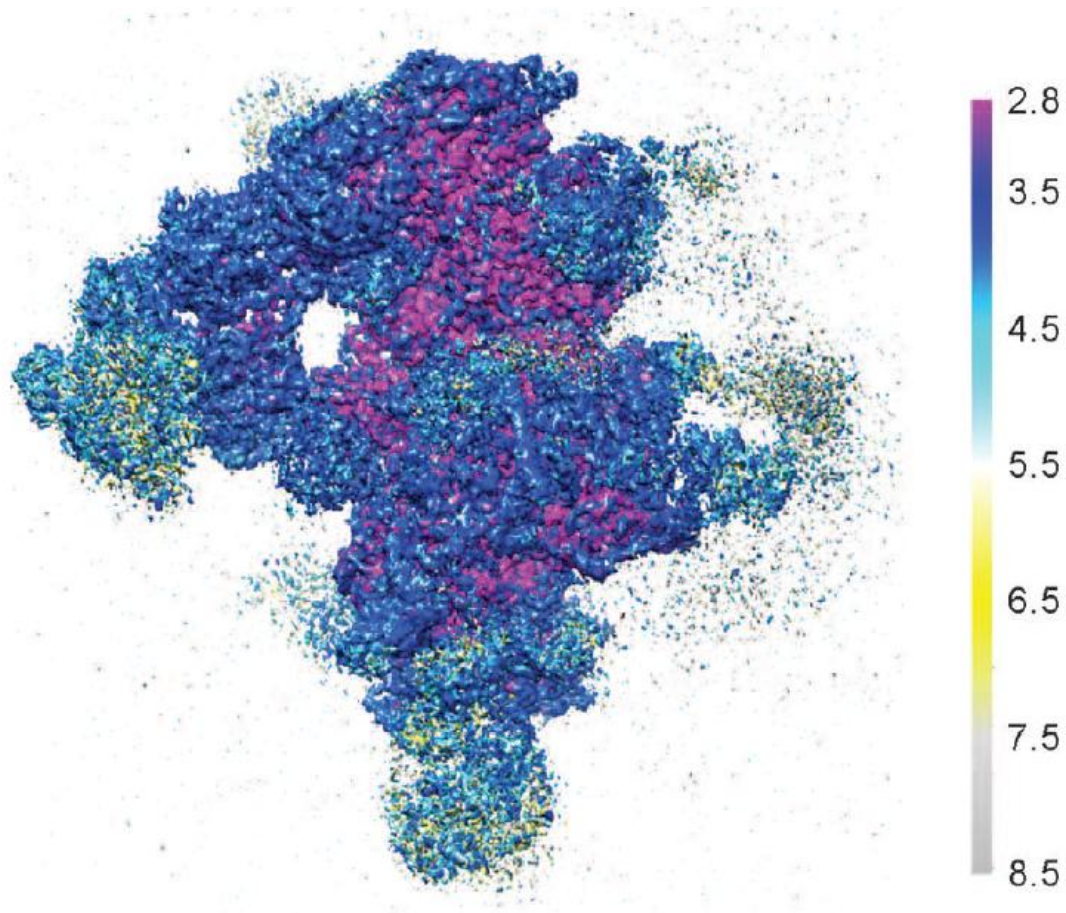


Figure 1.8: cryo-EM map for the yeast B<sup>act</sup> complex. Colors indicate resolution.

Adapted from Yan et al. 2016.

At round the same time, the structure of the spliceosomal C complex was also solved via Cryo-EM at an average resolution of 3.4 angstroms (Figure 1.9, Wan et al. 2016) and 3.8 angstroms by two different groups (Galej et al., 2016). This structure adds further support to the hypothesis that during this complex, the second step of splicing is poised to happen, but has not happened yet as the scissile phosphodiester bond has not been loaded onto the active site yet. What is worth noting, is that when comparing the C complex to the B<sup>act</sup>, while most proteins conform to identical conformations, the ribonuclease H (RNaseH)-like domain of Prp8 exhibits a large positional shift of up to 99 angstroms (Wan et al. 2016). Despite being shifted between



structures, the RNaseH-like domains align to themselves with near perfect registry. Thus it is proposed that the domain itself is ridged, but serves as a highly mobile element in the various spliceosomal complexes.

The structural analysis of the different spliceosomal complexes tremendously aided our understanding of the mechanistic details of the steps of spliceosomal assembly and continued work in this area will contribute to helping the field answer questions such as how splicing fidelity and efficiency is achieved.

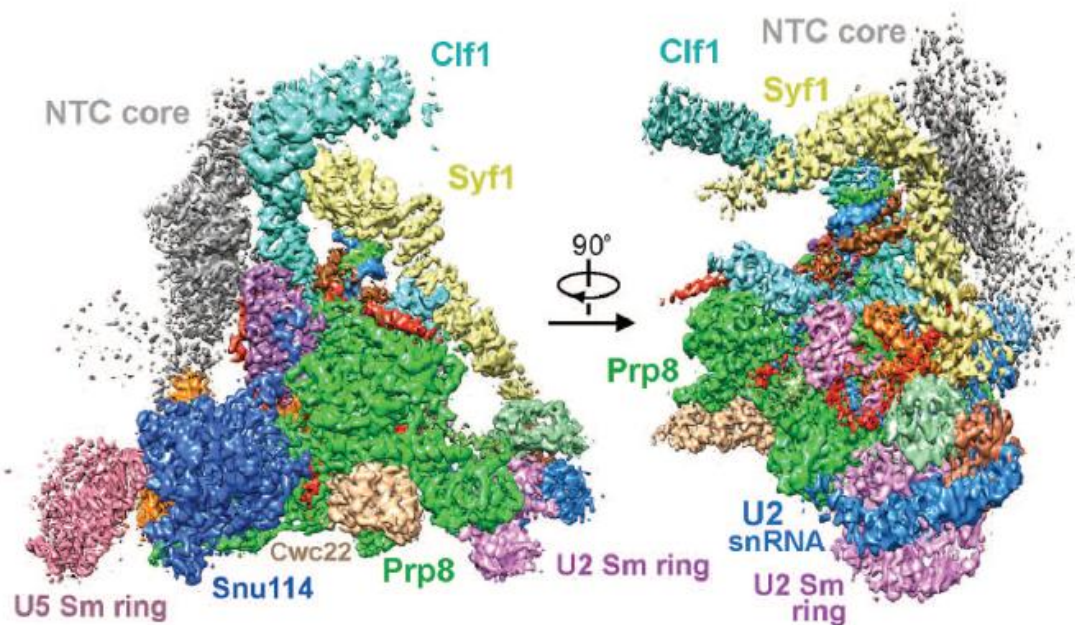


Figure 1.9: Cryo-EM map for the yeast C complex. Adapted from Wan et al. 2016.

#### 1.4: Splicing is an important and essential process in biology

Splicing is an important biological process and deeply involved in a wide area of topics ranging from gene expression regulation to protein diversity, to diseases. For example, the process of alternative splicing can generate additional protein variants from a single gene that contain many different introns. This phenomenon is especially telling when we compare humans, where there are about 20,000 genes present, to the

budding yeast *S. cerevisiae*, which has about 5000 genes. On the surface, it seems quite impossible that a 4-fold difference in gene numbers can generate the extreme complexities of humans compared to yeast. However, additional protein diversity is achievable via alternative splicing, generating many different mRNA isoforms from a single given gene (Nilsen and Graveley 2010). The alternative isoforms produced are also highly tissue dependent (Yeo et al. 2004), further increasing the diversity of proteins in humans. The increased protein diversity because of alternative splicing is essential in regulating tissue and organ development (Baralle and Giudice 2017). In addition to increasing protein diversity, splicing is also implicated in many diseases. For example, cystic fibrosis is a disease caused by the loss of function of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Polymorphism in intron 8 of this gene affects exon 9 splicing which leads to a nonfunctional protein, directly contributing to the disease (Faustino and Cooper 2003). For researchers involved in monitoring these changes in gene expression, the substantial number of mRNA variants produced through alternative splicing increases the difficulty by which they can be monitored.

### **1.5: Splicing is coupled to transcription**

Although it is possible to perform splicing *in vitro* using whole cell extract (Lin et al. 1985), suggesting that it is an independent process, we now know that splicing is connected to other cellular processes both temporally and spatially. Temporally, the process of splicing can happen as soon as the nascent RNA emerges from the polymerase, U1 and U2 deposition occur while transcription is still happening on the downstream portions of the RNA (Saldi et al. 2016). Spatially, the c-terminal domain (CTD) of RNA polymerase II is known to be associated with many splicing factors, such as the PRP19C complex (David et al. 2011). This coupling between splicing and other biological processes is difficult to model using *in vitro* studies, making them



somewhat less than ideal. To investigating splicing in its most native setting, *in vivo* studies are preferred, and a method that is able to accurately measure the splicing reaction *in vivo* is required.

### **1.6: Yeast as a model organism for studying splicing**

The baker's yeast *S. cerevisiae* is an excellent model organism used for studying splicing (Hossain and Johnson 2014). Being a highly tractable organism, genetic manipulation has been used with significant effect to study splicing in yeast. *S. cerevisiae* has a rather simple intronome, having lost most introns from its genome from its last eukaryotic common ancestor (Hooks, Delneri, and Griffiths-Jones 2014). Out of a total number of ~5000 genes, only about ~300 intron containing genes remain. Although baker's yeast has a simple intronome, the core spliceosomal machinery is highly conserved, so findings derived from studies in baker's yeast are often applicable to higher systems. Since many of the genes involved in gene expression regulation, including splicing, are essential, preventing studying by loss of function, the yeast model system provides an excellent alternative way of studying their mutants through temperature sensitive mutations. Finally, even though the number of intron containing genes are low in yeast and only make up ~5% of the genome, almost 40% of the transcriptome is made up of those intron containing genes (Ares, Grate, and Pauling 1999). Their over representation signifies the key role they play for the transcriptome, which is an excellent system to have for understanding the splicing process.

### **1.7: Historical methods used to monitor splicing**

Even though tremendous progress has been made in the field of splicing in the past decade, many important questions remain to be answered. For example, given the high degeneracy of splice sites, how does the spliceosome choose the “correct” site?

What are the cis and trans elements that affect splice site selection? What roles do cis and trans elements play in splice site fidelity? What roles do transiently associated factors to the spliceosome play in splice site selection and fidelity and what are the consequences of splicing mistakes? In order to probe these questions, we need a robust method to measure and quantify the changes in splicing caused by mutating these different factors. By studying mutants of these factors and observing the disruptions it causes in the splicing process, we can gain further understanding of the splicing process itself.

When splicing was initially discovered, it was studied on a per transcript basis using *in vitro* tools (Berget, Moore, and Sharp 1977) where the different products of splicing, typically of different lengths, were resolved through gel electrophoresis. This method is excellent at qualitatively determining the presence or absence of the products of different steps of splicing, but not very quantitative in terms of determining the relative ratios of the different products. Furthermore, the throughput of this method is very low, and is not suitable for looking at large numbers of different splicing targets.

Microarray hybridization techniques were then developed and provided a high-throughput, global examination of splicing profiles where probes against many exon-exon and intron-exon junctions are present on the array (Clark, Sugnet, and Ares 2002, Pleiss et al. 2007). Though relatively low cost for the amount of information gained per experiment, these microarray based approaches do suffer from several limitations. First of all, in order to construct the microarrays, prior knowledge of the sequences around the intron-exon and exon-exon boundaries must exist. This means the method isn't suitable for detection of novel events, such as activation of cryptic splice sites. Secondly, high background noise levels are typical owing to partial hybridization of off-target transcripts (Okoniewski and Miller 2006), this makes quantitative assessment difficult, especially when comparing between samples. Finally, since hybridization techniques

usually rely on fluorescence labeled targets, dynamic range is poor as signal can often be saturated on the high end. This masks the observable range of splicing defect different mutants can have on the same transcript.

### **1.8: Current methods used to monitor splicing**

With the rapid development of deep sequencing technology in the last decade or so, RNA-Seq had been quickly applied for use in field of splicing and has then become the standard used to monitor levels of splice isoforms. Because it does not rely on knowledge of existing genomic sequences, introns can be unambiguously annotated *de novo* at single nucleotide resolution from transcriptome data using computational techniques such as TopHat (Trapnell, Pachter, and Salzberg 2009). RNA-Seq also has no upper limit of detection, it is directly correlated to the number of reads obtained, and thus has an extremely high dynamic detection range.

One example of RNA-Seq's application in studying splicing is its use in the worm disease model of the retinitis pigmentosa (RP). RP is a group of genetically heterogeneous retinal diseases and a common cause of blindness. Surprisingly, 7 out of 24 of the autosomal dominant RP genes identified encode ubiquitous proteins essential for splicing (Rossmiller, Mao, and Lewin 2012). Why specific mutations in these broadly expressed, conserved pre-mRNA processing factors, PRPF3, PRPF4, PRPF6, PRPF8, PRPF31, SNRNP200/BRR2, and RP9, seem to only affect the eye is poorly understood, and understanding the molecular mechanisms for how splicing is involved in RP pathogenesis will impact therapeutic approaches for RP (Mordes et al. 2009). Traditionally, most efforts in trying to understand RP have been focused on a small set of genes and proteins that are known to be relevant to the disease. With RNA-Seq, genome-wide profiling of gene expression changes in the background of RP becomes possible, and the genetically tractable *C. elegans* which expresses all the adRP genes

with the exception of RP9, is used as a shortcut to understand why a mild reduction in the activity of genes required in all cells is critical only for a specific tissue. With the aid of RNAi to partially knock down adRP genes, it is observed that in developing worms, RNAi treated samples show an intron retention phenotype compared to wildtype (Rubio-Peña et al. 2015). Though this is an important finding and provided the authors with the hypothesis that inefficient splicing in the retina, where cells present a high transcriptional activity, causes a reduction in the mRNA levels and subsequently a deficit in the amount of retinal proteins produced leading to retinal degeneration, it is nevertheless unsatisfying that no transcript specific conclusions are made. It is of no surprise that mutations in the splicing machinery designed to remove introns will cause increased intron retention. What would be instrumental in progression towards therapeutics development for RP is finding out specifically which transcript's splice isoforms were altered to a detrimental degree, which this dataset was not able to find. For example, even at the depth of ~50 million reads per sample that the experiment was sequenced at, only 8 genes had enough depths to be quantitatively examined between the experiment and control samples and were found to not have changes in alternative splicing, despite ample RT-qPCR data suggesting PRPF8 depletion greatly alters alternative splicing (Tanackovic et al. 2011). This shows that despite RNA-Seq's ability to accurately quantify gene expression across samples, it performs poorly when it comes to sampling and quantifying splice isoforms (Cloonan et al. 2008).

The reason for RNA-Seq's poor ability to quantify splice isoforms is simple: reads that are informative of splice isoforms are rare by nature. Only reads that cross junctions, either between intron and exon, or between exon and exon, are informative of a transcript's splice isoform. This would mean that when using RNA-Seq to quantify transcript splice isoforms, only the small subset of junction spanning reads are useful, and all other reads are "wasted". Since junctions make up only a tiny fraction of the pre-

mRNA or mRNA, RNA-Seq reads that are splicing informative are rare, low count events (Figure 1.10). As an example, in *S. cerevisiae*, in a typical RNA-Seq experiment, only 0.3% of all reads map to splice informative junctions.

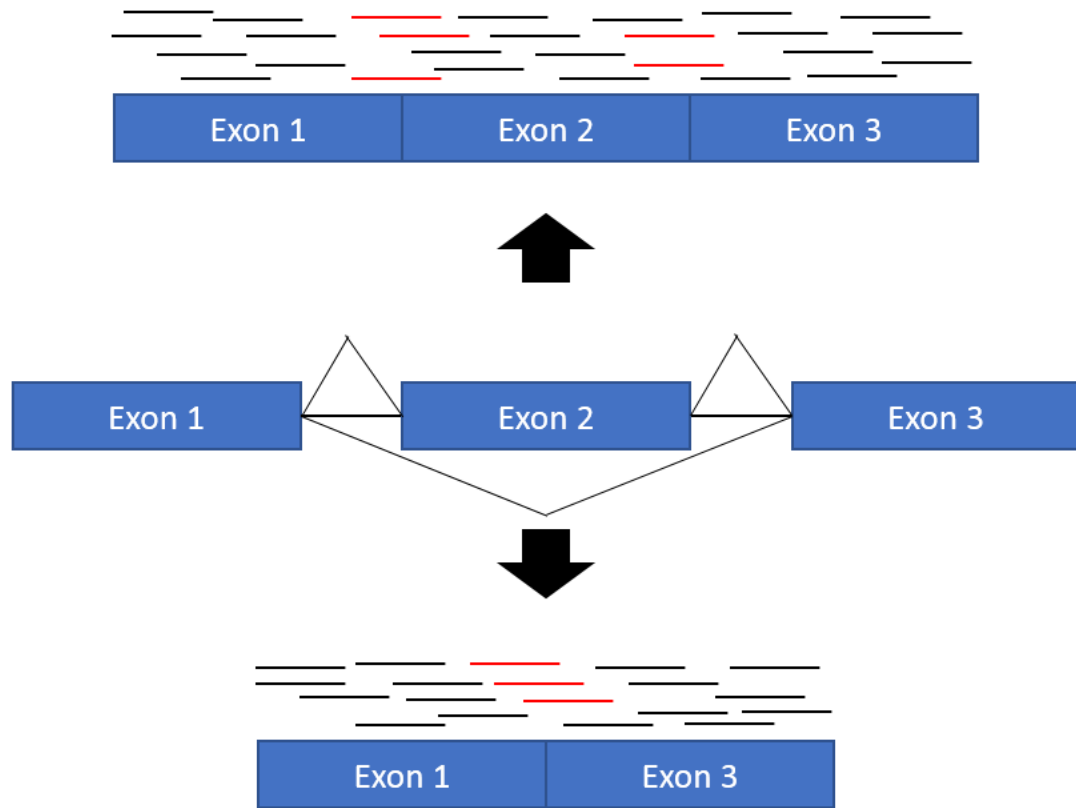


Figure 1.10: Few reads from RNA-Seq are splice informative. Representation of splice informative reads in RNA-Seq for an example alternatively splice transcript is shown.

Top panel represents canonically spliced product, bottom panel represents exon skipped product. Black lines represent mapped RNA-Seq reads that are not splice information, red lines represent reads that are splice informative.

While a powerful technique for examining and comparing between high frequency events, what is less well understood is the challenges RNA-Seq face when quantifying low frequency events. RNA-Seq is inherently noisier when counts for an event are low due to the fact that read counts model after a negative binomial

distribution (Law et al. 2014), which means that as count frequency gets lower, the variance becomes higher (Love, Huber, and Anders 2014). This becomes a real problem when one is trying to quantify different low count species in a given state as well as how they change in an altered state. Many have thought about the problem and have come up with different ways to deal with it. Mathematical methods have emerged to increase the confidence of low read count events. One such example, the mixture of isoforms (MISO) model, uses additional information from lengths of library inserts in paired-end data, recasts the analysis of splice isoforms as a Bayesian inference problem in order to obtain a higher confidence interval for estimates of splice isoform abundance than RNA-Seq without correction (Katz et al. 2010). However, even with proper mathematical corrections, RNA-Seq was still only able to accurately quantify splicing levels for ~55% introns in a ~140 million read mouse experiment (Herzel and Neugebauer 2015). This is particularly disappointing because ~140 million reads represent about half of the output of an Illumina NextSeq sequencing lane, and additional depth would mean that the experiment would quickly become cost prohibitive to perform, especially with the inclusion of multiple mutant conditions and replicates.

One way of tackling the problem of lowly quantitated species is to specifically block their degradation, thus achieving passive enrichment. For example, when Dbr1, an enzyme crucial for the degradation of lariat intermediates, is knocked out in the fission yeast *S. pombe*, lariat intermediates accumulate and can be subsequently sequenced by RNA-Seq (Taggart et al. 2012, Awan, Manfredo, and Pleiss 2013, Stepankiw et al. 2015). However, even with this method, only a mere 0.25% of mapped reads come from lariat intermediates if they are not actively enriched first (Bitton et al. 2014). Even though valuable information regarding 5'SS usage and branchpoint location is gained through the lariats identified, in general the per splice isoform counts are too low to be used in accurately quantitating splice isoform abundance vs. the

canonical isoform.

To overcome RNA-Seq's weakness with low frequency events, additional depth is the most crude and simple solution. However, even as the cost of RNA-Seq has come down dramatically over the last few years, increasing sequencing to the depth necessary to be able to precisely quantitate changes in splice isoforms is still cost prohibitive and folds higher than the depth that a typical experiment generates, especially when many mutants need to be compared. Rather, a method that specifically enriches reads to targets of interest is required.

### **1.9: Motivation for this dissertation**

In order to decipher the mechanistic details of the process of splicing, we are in need of a cost effective, accurate and precise method to globally monitor the products of both steps of splicing. Thus I have set out to develop one such method, Splice-Seq, which aims to utilize the depth gained through RNA-Sequencing and targeting that depth to specific regions of interest, namely the exon-exon and exon-intron junctions of intron containing transcripts. Splice-Seq can accurately quantify products of both chemical steps of splicing in a variety of conditions and mutant backgrounds at a low cost to the user. Though the technique is designed with the need of splicing in mind, it benefits from being easily modifiable into a wide variety of other targeted applications should it be desired.

## **Chapter 2: Splice-Seq as a novel targeted sequencing approach to study pre-mRNA splicing**

### **2.1 Abstract:**

Most eukaryotic genes have their protein coding sequences interrupted by non-coding introns, which must be removed from nascent pre-mRNAs by the spliceosome to generate a translatable mRNA. The past few decades have been marked by a significant increase in our appreciation for the central role that splicing plays in regulating eukaryotic gene expression. It is now clear that many organisms vastly expand their proteome by alternatively splicing pre-mRNAs to generate multiple protein isoforms from a single genetic locus, and that ever-increasing numbers of human diseases involve mis-regulation of this pathway. Nevertheless, the mechanisms by which this process is normally regulated, and can be mis-regulated with pathological consequences, remain poorly understood.

Next-generation sequencing (NGS) technologies have had a profound effect on our understanding of pre-mRNA splicing. By identifying the small subset of reads that span exon-exon junctions, implementation of NGS via RNA sequencing (RNA-Seq) has enabled the unambiguous detection of vast numbers of novel splice isoforms generated within a cell. Yet in spite of the power presented by this approach, it is less widely appreciated that the depth of sequencing necessary to quantitatively detect many splicing isoforms is significantly higher than most RNA-Seq experiments generate. Indeed, because of the error associated with low count numbers, the paucity of splicing-informative reads present in a standard RNA-Seq experiment significantly limits the ability to robustly assess quantitative changes in splicing patterns of specific transcripts in the background of different samples. As such, a deep understanding of the basic mechanisms by which splicing is regulated would benefit greatly from methods that enable higher resolution and precision detection of splicing states within cells.



Here I present the development and implementation of a novel sequencing method designed to harness the quantitative power of sequencing while focusing it on user-selected splice junctions of interest. This is accomplished through targeted reverse transcription where a pool of primers which anneal downstream of introns of interest are used instead of random priming. I demonstrate the ability of this approach to dramatically enrich the fraction of reads in a given sequencing experiment that are informative about splicing status, and in doing so enable a significant increase in the precision with which changes in splicing can be detected, all while decreasing experimental costs. Finally, I demonstrate the relative ease with which this approach can be adopted to diverse systems, facilitating a wide variety of experiments designed to understand the mechanistic underpinnings of splicing regulation.

## **2.2 Introduction:**

Most eukaryotic genes have their protein coding sequences interrupted by non-coding introns. Intron removal is catalyzed by the spliceosome, a complex and dynamic macromolecule comprised of 5 small nuclear RNAs (snRNAs) and hundreds of proteins. Spliceosomes assemble anew on every splicing substrate, recognizing sequence motifs within and around the defined pre-mRNA boundaries, enabling intron removal via two sequential transesterification reactions (Will and Luhrmann 2011). The past few decades have been marked by a significant increase in our appreciation for the central role that splicing plays in regulating eukaryotic gene expression. For example, using both spatial and temporal regulation, it is now clear that many organisms vastly expand their proteome by alternatively splicing pre-mRNAs to generate multiple protein isoforms from a single genetic locus (Kelemen et al. 2013). Nevertheless, in spite of its central importance in the gene expression pathway, the mechanisms by which this process is regulated remain poorly understood.

Over the past decade, next-generation sequencing (NGS) technologies have had a profound effect on nearly all facets of modern biology, including pre-mRNA splicing. NGS studies designed to understand the genetic underpinnings of various human diseases highlight the critical role of splicing in human biology: it is now estimated that as many as 50% of all genetically heritable human diseases impact the splicing pathway (Ward and Cooper 2010), either through mutations in the spliceosomal machinery itself, or through mutations within individual transcripts that alter their splicing. Simultaneously, implementation of NGS via RNA sequencing (RNA-Seq) has greatly expanded our understanding of the variety of spliced isoforms that can be generated within a cell. Identification of the small subset of reads that span exon-exon junctions within transcripts has enabled unambiguous detection of vast numbers of novel splice isoforms in scores of organisms (Trapnell, Pachter, and Salzberg 2009). Yet in spite of

the power presented by this approach, it is less widely appreciated that the depth of sequencing necessary to quantitatively detect many splicing isoforms is significantly higher than most experiments generate. Indeed, because of the error associated with low count numbers in RNA-Seq experiments (Katz et al. 2010), the paucity of splicing-informative reads present in a standard RNA-Seq experiment not only limits the ability to detect minor splice isoforms within a given sample, but also the ability to robustly assess quantitative changes in splicing patterns of specific transcripts in the background of different samples. A deep understanding of the basic mechanisms by which splicing is regulated, and the pathological consequences of its mis-regulation, will benefit greatly from methods that enable higher resolution and precision detection of splicing states within cells.

Here we present the development and implementation of a cost-effective, targeted sequencing approach that enables high precision detection of the genome-wide products of both chemical steps of pre-mRNA splicing. Though the technique is designed with the need of splicing in mind, it benefits from being easily modifiable into a wide variety of other targeted applications should it be desired.

### 2.3 Results:

In order to better detect and understand genome-wide changes in pre-mRNA splicing, we designed and implemented a method, herein referred to as Splice-Seq, that enables targeted enrichment of sequencing reads at user-selected splice junctions. Building upon the historically validated use of primer extension as a tool for assessing splicing status, Splice-Seq was developed as a high-throughput variation of this approach coupled to sequencing as a tool for digital quantitation, a schematized depiction of which is presented in Figure 2.1A. As an initial test of this approach we designed an experiment to monitor pre-mRNA splicing in the budding yeast *Saccharomyces cerevisiae*, an organism for which the core spliceosomal machinery is highly conserved with multicellular organisms, yet which harbors a highly reduced set of introns within its genome, simplifying the process of genome-wide analyses.

*Primer Design:* For each of the 309 annotated introns within the budding yeast genome, a reverse transcription primer was designed within the first 50 nucleotides downstream of the intron. Targeting to this region ensured that short-read sequencing of the products generated from reverse transcription with these primers would cross the upstream exon-exon or exon-intron boundaries, enabling determination of the splice isoform. Primers were designed using OligoWiz, a program initially developed for microarray probe design, but which enables the selection of primer sequences optimized for target specificity relative to a designated genomic background (Wernersson, Juncker, and Nielsen 2007). To the 5' end of each of these primers was appended two additional sequence elements: an 8 nucleotide random region which allows for the detection and removal of amplification artifacts arising from library preparation (Shiroguchi et al. 2012); and the P500 region of the Illumina sequencing primer to enable the sequencing of the reverse transcription products. Each of these primers was individually synthesized, the full sequences of which are provided in the supplemental

materials for Xu et al. 2018, in preparation.

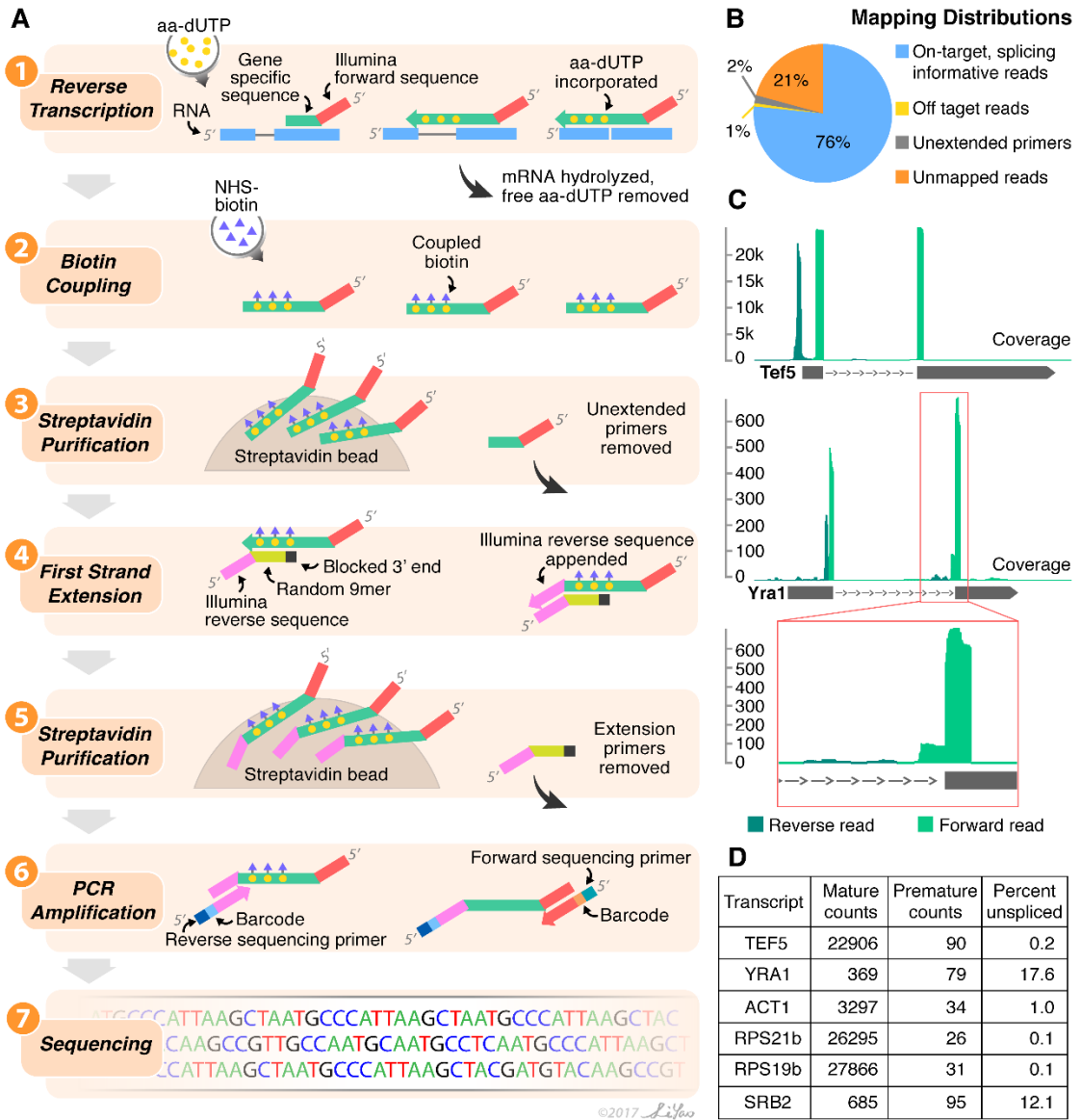


Figure 2.1: Splice-Seq can quantitatively monitor splice isoforms. (A) Schematic showing the workflow of Splice-Seq. (B) Splice-Seq mapping percentages for *S. cerevisiae*. (C) Coverage visualization for Splice-Seq mapping to example genes Tef5 and Yra1. (D) Five example genes and their isoform numerical data from Splice-Seq are shown.

*Targeted cDNA synthesis:* Beginning with total cellular RNA, cDNA was synthesized in a reaction containing a total of 1 µg of primer, consisting of equimolar amounts of each of the 309 individual primers. Reactions were performed under generally standard conditions (see Methods for additional details) but in the presence of the nucleotide analog aminoallyl-dUTP (aa-dUTP) in order to facilitate recovery of cDNA products. In order to increase primer specificity and decrease off-target annealing, reactions were incubated at 47 °C. After removal of the RNA and unincorporated aa-dUTPs, the extended cDNA molecules were biotinylated through the incorporated aminoallyl-dU residues by reaction with NHS-biotin. The biotinylated cDNA was then purified using streptavidin beads, enabling removal of the excess of unextended primers.

*First strand extension:* Conversion of the products of these reverse transcription reactions into sequence-able material was accomplished using a modified version of second strand synthesis that enabled preservation of the 3' end of the cDNA products. We employed a strand extension technique that included as a 'second strand primer' a dN9 template oligo with a blocked 3' end that was incapable of being extended by Klenow (Hexanediol Spacer from Integrated DNA Technologies), and to which the Illumina P700 sequence was appended on its 5' end. Whereas these molecules can anneal via the random region to virtually any portion of the generated cDNA products, they cannot be extended by Klenow by virtue of their blocked 3' ends. By contrast, when these primers anneal to the 3' terminus of the cDNA product, Klenow can extend the first strand products by appending the Illumina P700 sequence. After re-purification of the extended first strand products and removal of the 'second strand primers' by streptavidin purification, the products of these first strand extension reactions were amplified using primers containing Illumina sequences that enabled bar-coded library preparation.

*Splice-Seq determines the splicing status of intron containing transcripts with high precision:*

As an initial implementation of Splice-Seq, we designed a sequencing experiment with a targeted depth of approximately 5 million reads. While this level of sequencing corresponds to just a small fraction (1-2%) of the capacity of a single run on an Illumina NextSeq, on the basis of total expression levels derived from published RNA-Seq datasets we estimated that this would provide sufficient depth to sample the majority of splice junctions (Nowrousian 2013). Libraries were generated according to the protocol described in Figure 1A (see also Methods for details), and subsequently sequenced on an Illumina NextSeq. For reasons that will be described more completely later, the sequencing was performed such that 60 nucleotides were sequenced from the forward end of our library (corresponding to the splice junction region) and 15 nucleotides were sequenced from the reverse end (corresponding to the terminus of the cDNA). Paired-end reads were aligned using the Star Aligner with typical settings (see Methods).

To assess the effectiveness of Splice-Seq for enriching splicing-informative reads, we first asked about the overall alignments generated by this experiment. As seen in Figure 1B, over 75% of the reads were splicing informative, meaning that they were derived from one of the 309 designed primers, and reflected extension of a cDNA product across the upstream splice junction. By contrast, only a small fraction (~1%) of the reads aligned to regions not targeted by one of the designed primers, in most cases reflecting cross-hybridization with a poorly matched region within a highly expressed RNA (Figure 2.2). Similarly, a small fraction (~2%) of the reads reflected unextended primers, wherein the synthesized primer sequences appeared directly appended to the reverse Illumina primer. The remaining ~20% of the reads reflected a combination of:

on-target, splicing informative reads that nevertheless contained too many mismatches to allow for proper alignment; and low information sequences that were likely amplification or sequencing artifacts. Taken together these data confirmed the ability of this approach to efficiently target sequencing reads to regions of interest.

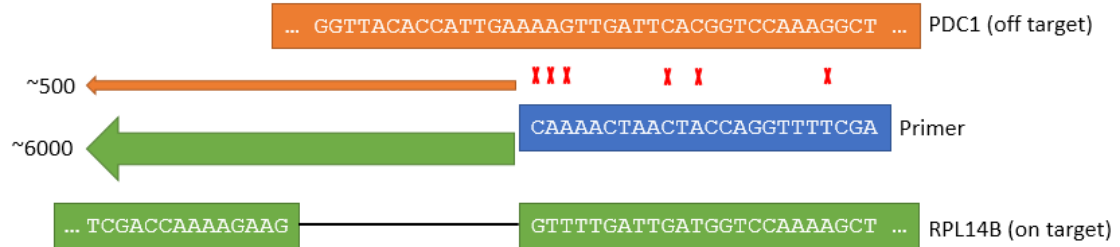


Figure 2.2: A small percentage of reads are mapped to off target transcripts due to partial matches in the primer sequence. The primer (blue) has partial mismatches to the off target transcript PDC1 (orange) and resulted in ~500 reads of the off target cDNA synthesized. The on target transcript RPL14B (green) has ~6000 reads.

In order to assess the ability of this approach to quantify splicing status, we examined the reads that were aligned to individual intron-containing genes. For the vast majority of genes, the data were consistent with the idea that levels of spliced mRNA are higher than unspliced mRNA in the steady state pool, in that most reads span the exon-exon boundary rather than reading from the downstream exon into the intron. Figure 1C shows examples of the read coverages associated with two genes, Tef5 and Yra1, and Figure 1D shows the quantitation for several additional genes (the full data are available in the supplemental materials of Xu et al. 2018 once published). Importantly, whereas the Tef5 transcript showed a high partitioning towards spliced mRNA, where tens of thousands of reads corresponded to the mature isoform but just under 100 reflected the presence of pre-mRNA, a much higher fraction of unspliced mRNA was detected for Yra1 (see lower inset of Figure 1C), a transcript that has previously been shown to be auto-regulated through its slow rate of pre-mRNA splicing.



Indeed, the steady-state splicing efficiencies determined from these data were largely consistent with previously reported values (Pleiss et al. 2007).

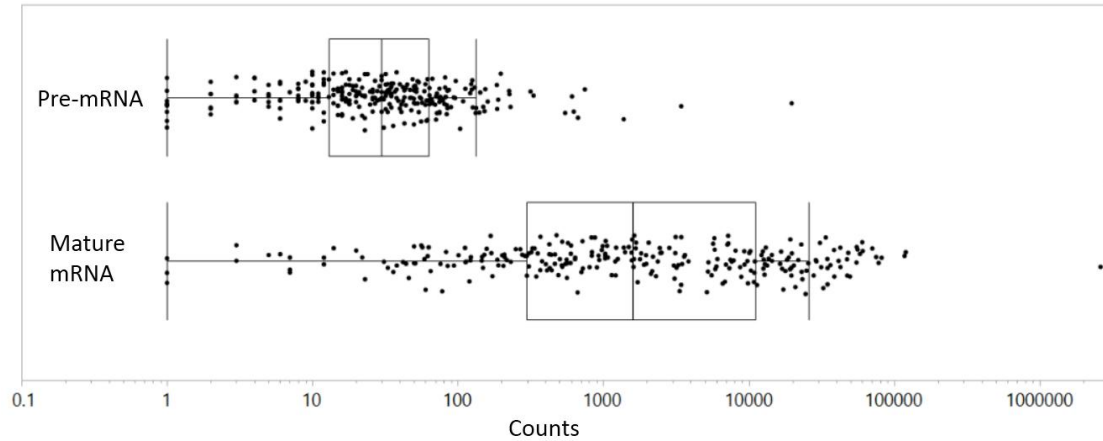


Figure 2.3: Read count distributions for both premature mRNA and mature mRNA are high in a 5 million read Splice-Seq experiment. Count distributions for both species are shown. Box covers 25<sup>th</sup> – 75<sup>th</sup> percentiles. Line inside box represents the mean.

Given the strong enrichment of reads targeted to the splice junctions in this initial experiment, we sought to examine the precision with which Splice-Seq measured splicing efficiency, particularly in comparison with RNA-Seq. Because robust numbers of reads were generated for both the spliced and unspliced isoforms of nearly every intron-containing gene (Figure 2.3), we expected that Splice-Seq would not be subject to the same level of sampling-related noise apparent in most RNA-Seq datasets, and would therefore yield more precise measurements of splicing efficiency. To test this, replicate Splice-Seq libraries were generated from a wild type sample of RNA using the previously described protocol, while additional replicate RNA-Seq libraries were generated using NEB’s RNA sequencing kit (following manufacturer’s protocol). Whereas the Splice-Seq libraries were again targeted for ~5M reads of sequencing each, the RNA-Seq libraries were targeted for ~30M reads each, a level more commonly used in RNA-Seq experiments (Sims et al. 2014, Conesa et al. 2016). For each intron-

containing gene we then calculated the percentage of total reads mapping to the intron that reflected unspliced mRNA. As seen in Figure 2.4A, Splice-Seq generated highly reproducible values for transcripts across a wide spectrum of relative splicing efficiencies ( $R^2=0.96$ ), consistent with the robust count numbers generated even for lowly expressed transcripts within this small dataset. Importantly, while we did not filter these data for a minimum read number, splicing efficiencies could be calculated in both replicate libraries for over 90% of the intron-containing genes, meaning that at least one read was detected for both splicing isoforms within both libraries. As an initial, direct comparison, the RNA-Seq libraries were downsampled to 5M reads apiece (the same size as the Splice-Seq libraries) with the percent unspliced again calculated for every intron-containing gene. Two important conclusions are apparent from these data, as shown in Figure 2.4B. First, splicing efficiencies could not be calculated for nearly 20% of the intron-containing genes because at least one isoform failed to be detected in at least one of the samples, reflecting the relatively low frequency at which short-read sequences traverse splice junctions in a traditional RNA-Seq experiment. Second, for the subset of intron-containing genes for which splicing efficiencies could be determined, the reproducibility was significantly worse than Splice-Seq ( $R^2=0.59$ ). When considering the full, 30M read RNA-Seq datasets, as seen in Figure 2.4C, the number of intron-containing genes for which splicing efficiency could be calculated returned to a level equivalent to Splice-Seq, yet the reproducibility continued to lag behind ( $R^2=0.86$ ) even with the significant increase in sequencing depth (and associated cost). To better understand the level of enrichment that Splice-Seq provided over RNA-Seq in these experiments, the number of exon-exon reads generated for each intron-containing gene was determined for each method, normalized to the total number of reads generated. As seen in Figure 2.4D, these data suggest a median enrichment of over 200-fold.

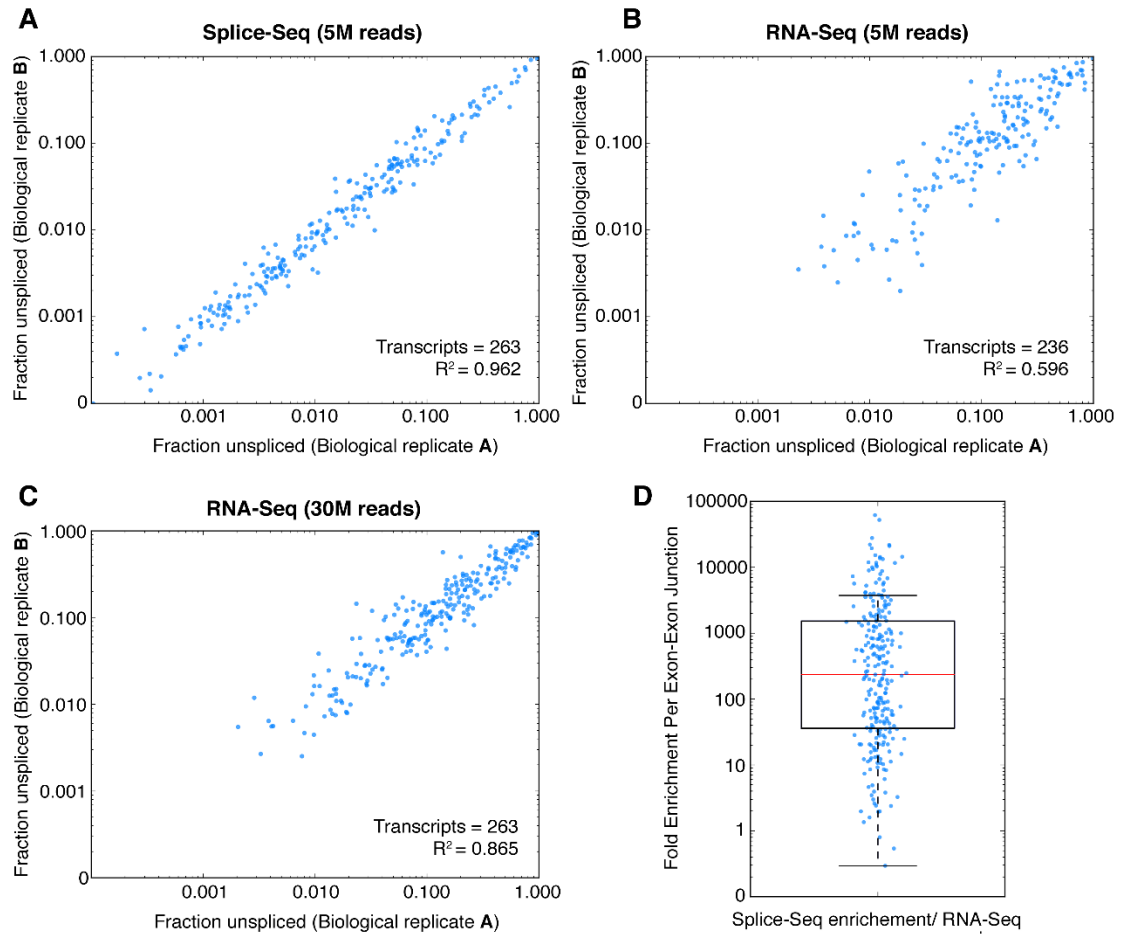


Figure 2.4: Splice-Seq can more precisely quantify splicing efficiencies than RNA-Seq using less depth. Fraction unspliced (premature counts/(premature counts + mature counts)) is calculated for wildtype biological replicates of BY4741 strain. Biological replicates are plotted on log log correlation plots shown for a 5M read Splice-Seq experiment (A), 5M read RNA-Seq experiment (B), and 30M read RNA-Seq experiment (C). Exon-exon count enrichment on a per transcript basis is plotted as a scatter plot in (D).

*Splice-Seq robustly quantifies changes in splicing efficiency.*

While understanding the splicing efficiency of transcripts in a particular state is important, for many researchers the critical metric to understand is how splicing efficiency is altered with changes in developmental or genetic condition. To confirm that Splice-Seq could robustly quantify changes in splicing efficiency between two different samples, we took advantage of a budding yeast strain containing a mutation in the well characterized spliceosomal factor Prp2 (Kim and Lin 1996). Prp2 is a helicase necessary for catalyzing a structural rearrangement in the spliceosome required for the first chemical step, and while strains harboring the *prp2-1* mutation are able to grow with nearly wild type efficiency at 25 °C, they are unable to support growth at 37 °C (Kim et al. 1999). We have previously shown that genome-wide defects in pre-mRNA splicing are apparent even after short times exposed to elevated temperature (Pleiss et al. 2007). To determine the capacity of Splice-Seq to detect changes in splicing efficiency associated with this mutant, we harvested RNA from matched wild type and *prp2-1* mutant strains, each of which had been shifted to the non-permissive temperature for 15 minutes. Figure 2.5A shows how splice efficiencies for a few example genes change in *prp2-1* with respect to wildtype under non-permissive temperatures. As before, we generated replicate Splice-Seq libraries from each of these strains as well as replicate RNA-Seq libraries, then sequenced each of them to generate ~5M and ~30M read datasets, respectively. As expected, Splice-Seq robustly and reproducibly ( $R^2=0.92$ ) measured the changes in splicing efficiency across a wide dynamic range, as seen in Figure 2.5B. Similarly, as previously demonstrated RNA-Seq showed far lower reproducibility at equal (5M) read depths ( $R^2=0.67$ , Figure 2.5C) and even at significantly elevated read depths ( $R^2=0.82$ , Figure 2.5D).

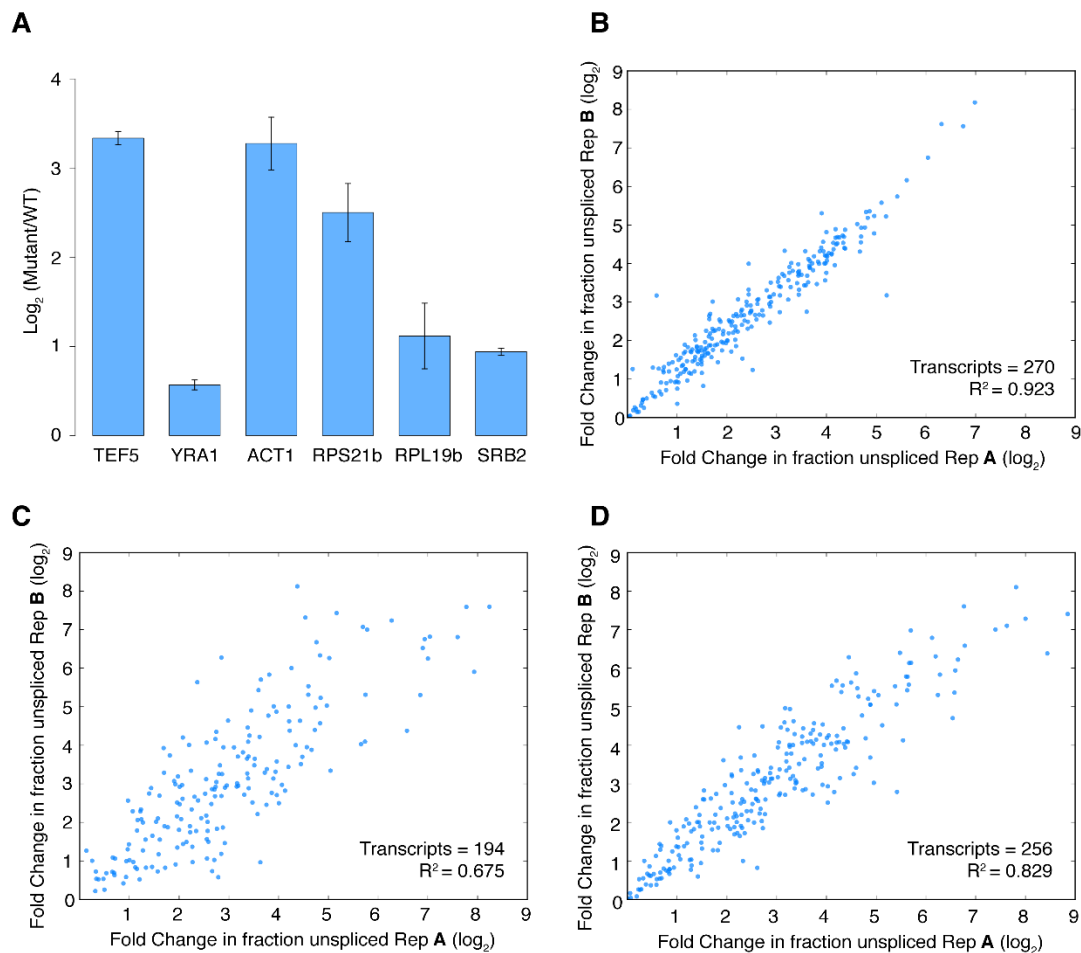


Figure 2.5: At the non-permissive temperature, *prp2-1*'s splicing defect can be precisely captured by Splice-Seq. (A) Change in fraction unspliced when shifted to the non-permissive temperature is shown for 6 example transcripts. Change in fraction unspliced for biological replicates are plotted against each other in panels (B), (C) and (D) for Splice-Seq at 5M reads, RNA-Seq at 5M reads and RNA-Seq at 30M reads respectively.

#### *Splice-Seq can identify and differentiate between splicing intermediates*

Our initial experimental design for second strand synthesis anticipated the possibility that this approach might enable detection of the locations of reverse

transcription stops. In its early use for studying pre-mRNA splicing, primer extension assays were often used to identify the lariat intermediate species that are generated after the first but prior to the second chemical steps of splicing (see Figure 2.7A) on the basis of a strong reverse transcription stop at the branched adenosine in the lariat intermediate (Coombes and Boeke 2005). Because the locations of introns within budding yeast genes show a strong positional bias towards the extreme 5' end, we imagined that most reverse transcription products generated in these experiments would extend to the natural 5' end of the mRNAs. In fact, as seen in the coverage map of the *Tef5* transcript in Figure 1C, a large number of reverse reads (those derived from the 3' end of the cDNA products) mapped to a region near the expected transcription start site (TSS) for this gene. While a range of similar but distinct global TSSs have been described in budding yeast using a variety of techniques (Pelechano et al. 2014, Booth et al. 2016, Zhang and Dietrich 2005), a comparison of the reverse transcription stops generated in our experiment with the predicted TSSs from these different studies all showed strong correlations (see Figure 2.6), consistent with the capacity of Splice-Seq to reveal the locations of reverse transcription termination.

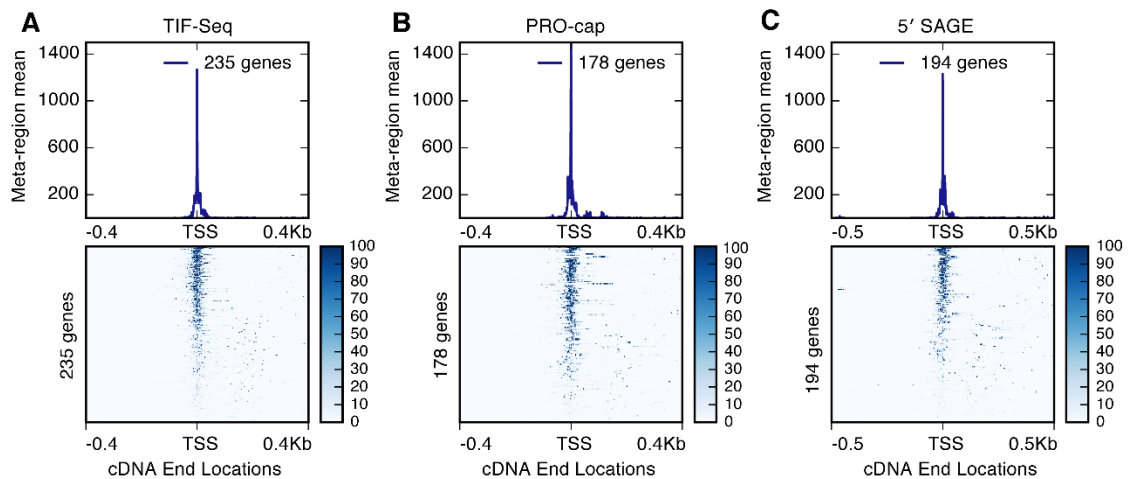


Figure 2.6: TSS mapped by Splice-Seq correlates well with existing datasets. TSS from Splice-Seq is mapped against TIF-Seq, Pro-Cap and 5' SAGE in panels (A), (B) and (C) respectively. Top panels show average cDNA end position centered around the annotated TSS from the indicated mapping methods. Bottom panels visualizes read count density normalized to 100 for all transcripts containing data on cDNA end positions centered at the annotated TSS from the indicated mapping methods.

Having demonstrated the capacity to map reverse transcription stop sites, we asked whether data from Splice-Seq could be used to differentiate between unspliced reads where reverse transcription proceeded beyond the branchpoint adenosine (presumably reflecting unspliced pre-mRNA which had not yet undergone the first chemical step of splicing) and those where reverse transcription stopped at the branchpoint adenosine (presumably reflecting the lariat intermediate product of the first chemical step). For each nucleotide within each intron-containing gene, we determined the fraction of cDNA molecules that included that position, the results of which are shown in Figure 2.7B. The upper portion of this figure shows the composite behavior as seen across all 254 transcripts detected in this experiment, whereas the lower portion shows a false-colored representation of each of the individual transcripts. When the data

were considered from right to left (that is, as the reverse transcription event moved from the 3' splice site towards the 5' splice site), a gradual decrease in read density was apparent, presumably reflecting the background rate of reverse transcription termination, mostly likely owing to poor processivity of the enzyme or random cleavage events in the underlying RNA. By contrast, a sharp decrease in read density was apparent precisely at the location of the annotated branchpoint, consistent with a strong stop in reverse transcription at this location. Interestingly, comparison of the profiles for individual transcripts revealed a wide variety of phenotypes: some transcripts showed relatively low fractions terminating at the branchpoint, consistent with a large population of unspliced mRNAs prior to the first chemical step, while others showed high fractions of lariat intermediate.



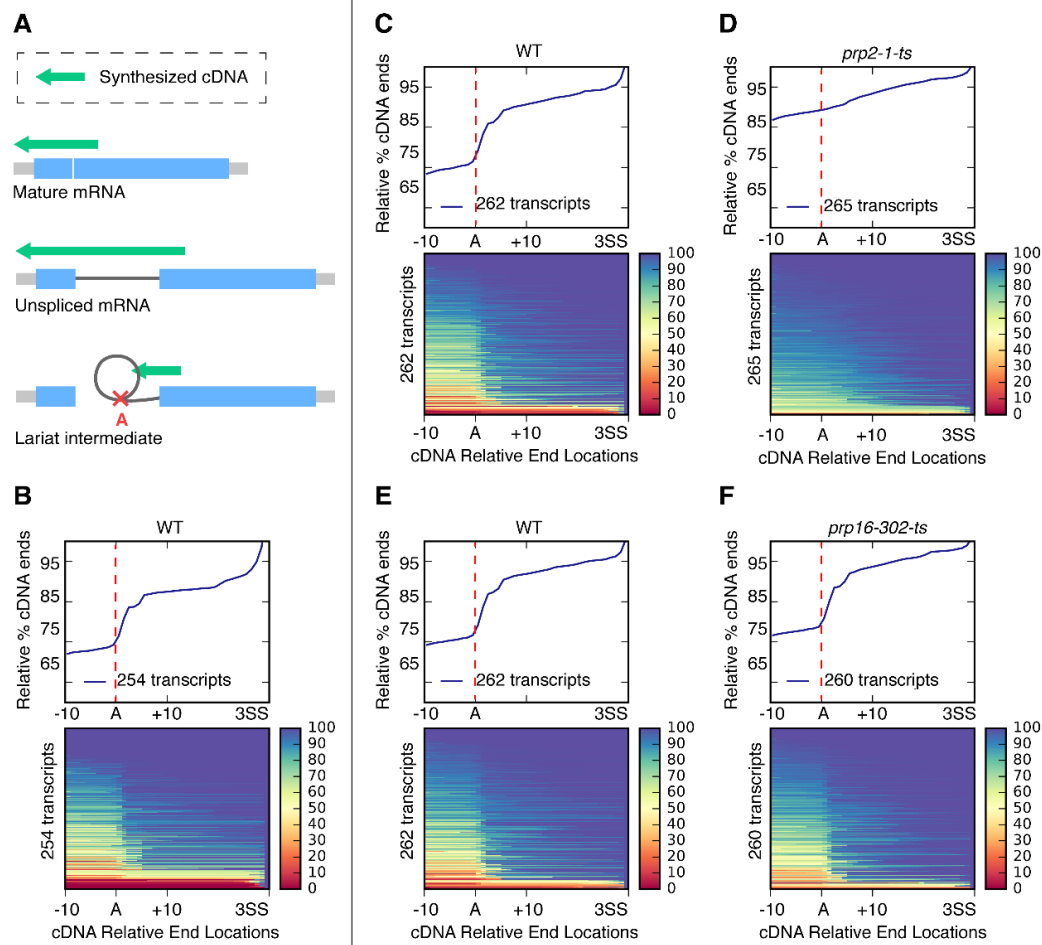


Figure 2.7: Splice-Seq is able to detect branchpoint accumulation in different splicing mutants. (A) Visualization of different end positions for termination of reverse transcriptase. (B) through (F) (B: wildtype under normal growth condition, C and D: wildtype and *prp2-1*, E and F: wildtype and *prp16-302*) Top panels: percentage of intronic cDNA reads terminating at each location centering at -10 to +10 bp of the annotated branchpoint A through the 3' splice site for each transcript. Normalized to 100% at the 3' splice site. Bottom panels: graphical representation of the intronic percentage of cDNA reads terminating at the same locations. Note: the apparent minor but sharp decrease located ~3 bp downstream of the annotated branchpoint A comes from a data processing artefact related to first strand extension primer having sequence overlap with the reverse Illumina sequence and is not biologically relevant.

To confirm that the reverse transcription stops we detected were indeed a result of the branched lariat structures, we returned to our mutant alleles. Using *in vitro* splicing systems, others have previously demonstrated that extracts from *prp2-1* strains are unable to catalyze the first chemical step of splicing: formation of the lariat intermediate (Silverman et al. 2004). As such, we performed a similar analysis on the libraries derived from our strain harboring the *prp2-1* mutation and its matched wild type. As seen in Figure 2.7C and D, whereas the strong reverse transcription stop remained apparent in the wild type sample, in the *prp2-1* sample there was little if any detectable lariat intermediate on either a global or transcript-level basis, consistent with the absence of lariat intermediate products in this sample. To augment this result, we designed a complementary experiment: whereas Prp2 is a helicase that functions immediately prior to the first chemical step, Prp16 is a helicase that functions after the first chemical step but prior to the second, as such we expected that mutations impairing Prp16 function would lead to an accumulation of lariat intermediates genome-wide with a concomitant decrease in the amount of fully unspliced pre-mRNA (Hogg et al. 2014). Splice-Seq libraries were generated from RNA from a strain harboring a *prp16-302* mutation along with a matched wild type strain after both had been shifted to the non-permissive temperature. Surprisingly, while accumulation of pre-mRNA was detected genome-wide in the *prp16-302* sample (see Figure 2.8A C and D), as seen in Figure 2.7F, the relative proportion of lariat intermediate present was largely unchanged, consistent with the idea that both the first and second chemical steps were defective. Although the time spent at the non-permissive temperature was relatively short for these samples, we imagined that inactivation of Prp16 could ultimately lead to a general defect in splicing if sub-stoichiometric components of the spliceosome became sequestered with stalled *prp16-302*-associated spliceosomes. We therefore generated additional

Splice-Seq libraries from the *prp16-302* strain shifted to the non-permissive temperature for a variety of very short times. Remarkably, while these samples clearly revealed time-dependent increases in global pre-mRNA levels (Figure 2.8C) and global lariat intermediate levels (Figure 2.8D), the relative amount of fully unspliced pre-mRNA versus lariat intermediate was nearly identical at all times, as shown in Figure 2.8B.

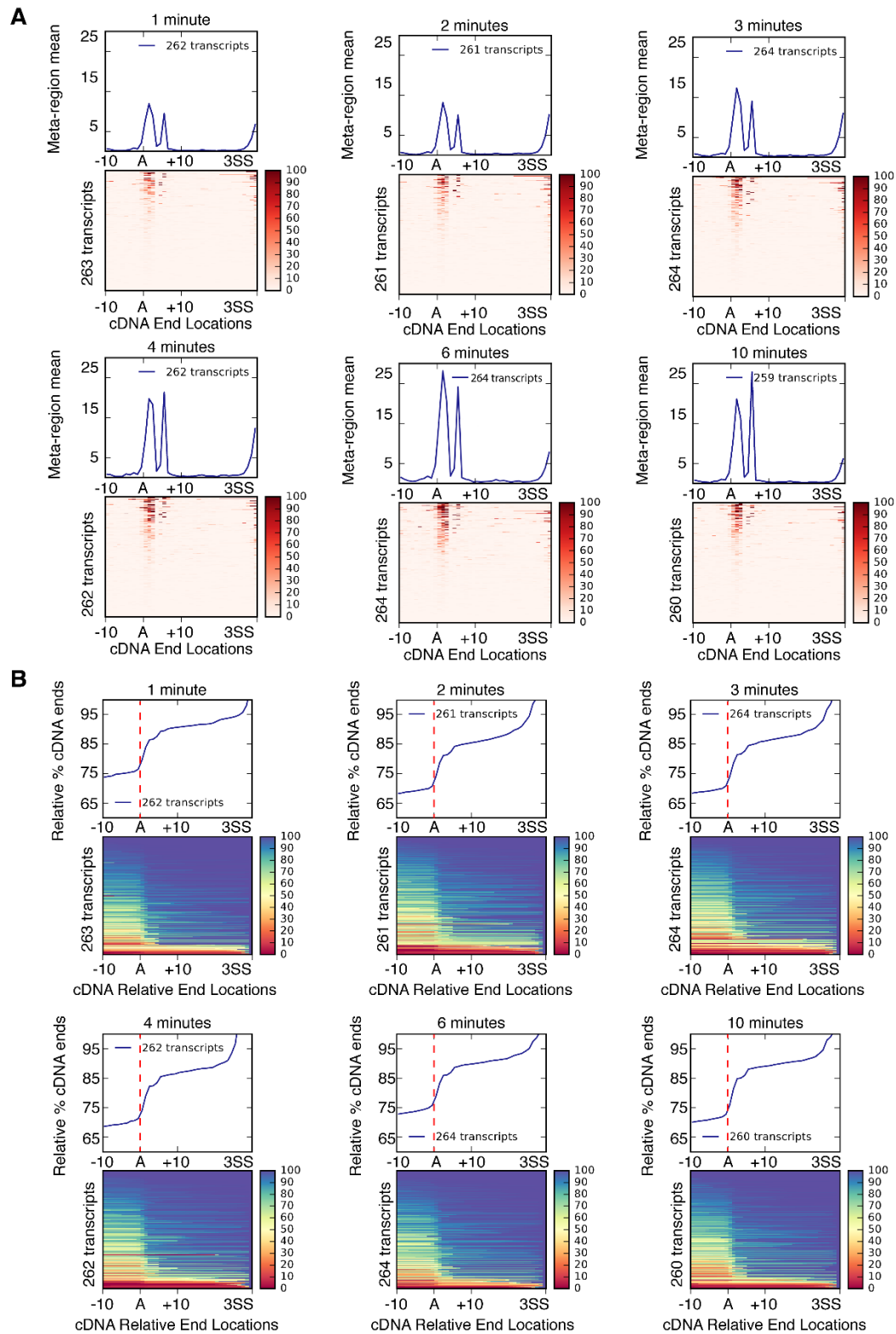


Figure 2.8: Continued on next page.

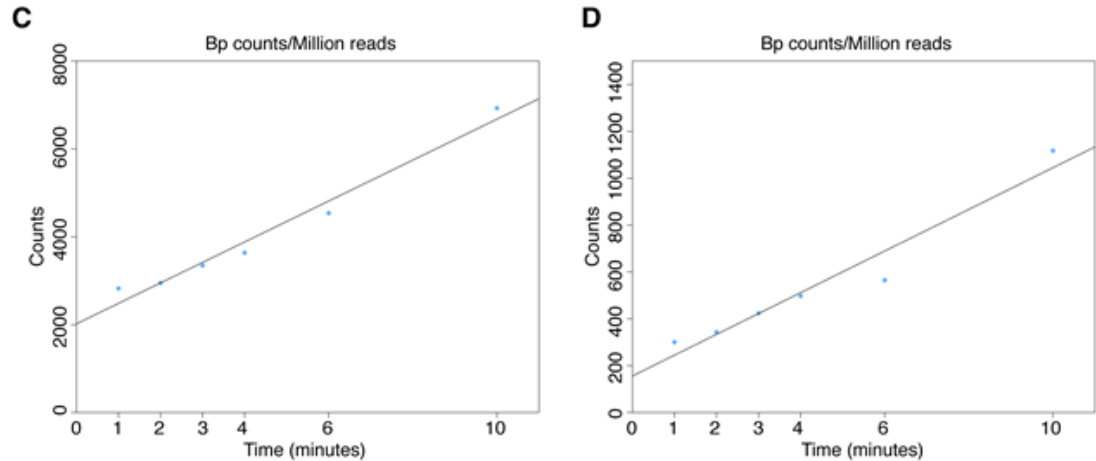


Figure 2.8: Continued from previous page. Time dependent lariat intermediate and pre-mRNA accumulation is observed for *prp16-302* but relative levels stay constant. (A) Top panels: Average counts for cDNAs terminating at positions centered at -10 to +10 bp of the annotated branchpoint A through the 3' splice site for each transcript. Bottom panels: Graphical representation for counts at each basepair location centering at -10 to +10 bp of the annotated branchpoint A through the 3' splice site for each transcript. (B) Top panels: percentage of intronic cDNA reads terminating at each location centering at -10 to +10 bp of the annotated branchpoint A through the 3' splice site for each transcript. Normalized to 100% at the 3' splice site. Bottom panels: graphical representation of the intronic percentage of cDNA reads terminating at each location centering at -10 to +10 bp of the annotated branchpoint A through the 3' splice site for each transcript. Normalized to 100% at the 3' splice site. (C) and (D) Counts of cDNAs terminating around the branchpoint A (+/- 4bp), or anywhere within the intron, respectively, is plotted over the time course. Normalized to sample read depth.

*An efficient method for generating complex pools of Splice-Seq primers.*

Although the above work demonstrated the capacity of Splice-Seq to robustly monitor global splicing status, the utility of the approach as described would be limited to those organisms for which it was economically feasible to synthesize the full complement of needed primers. While it would quickly become cost prohibitive to individually synthesize the thousands of discrete oligonucleotides necessary undertake experiments in an organism with complex splicing, many commercial sources exist today wherein tens or even hundreds of thousands of oligonucleotide sequences can be generated in a complex mixture, often utilizing microarray synthesis technologies. Although the quantity of any given oligonucleotide generated by these methods is vanishingly small, we designed a protocol by which these could be used as a template to generate microgram quantities of single-stranded primers, a schematic of which is shown in Figure 2.9A (see also Methods for details). Briefly, to the 3' end of each of the primer sequences presented in the supplemental table (available online once this work is published at Xu et al. 2018), we appended two additional motifs. First was a SapI restriction site, designed such that cleavage with SapI would generate the precise 3' end of the primers shown in the mentioned supplemental table. Second was a constant sequence: PCR reactions were then performed using a reverse amplification primer complementary to this constant sequence along with an Illumina forward primer (present on the 5' end of each of the designed oligos in the supplemental table). For reasons described below, the forward Illumina primer in these reactions was synthesized with a three carbon block on its 5' end, while the reverse amplification primer was synthesized with a biotin on its 5' terminus. After amplification with these primers and digestion with SapI, treatment with Lambda Exonuclease preferentially degraded those strands with a 5' phosphate generated by SapI digestion. The desired single stranded oligonucleotides were then purified by depleting all biotin-containing species, affecting

removal of both the single stranded reverse primer regions as well as any double stranded molecules that failed to undergo SapI digestion. As seen in Figure 2.9B, this approach allowed for robust generation of single stranded, pooled primers. To demonstrate the activity of these primers, they were used to generate replicate Splice-Seq libraries from the same wild type RNA we previously used. As before, the global splicing efficiencies determined from these replicate experiments were very well correlated with one another and importantly, a comparison of data generated with the different sources of primers also show a very high correlation ( $R^2=0.91$ , Figure 2.9C).

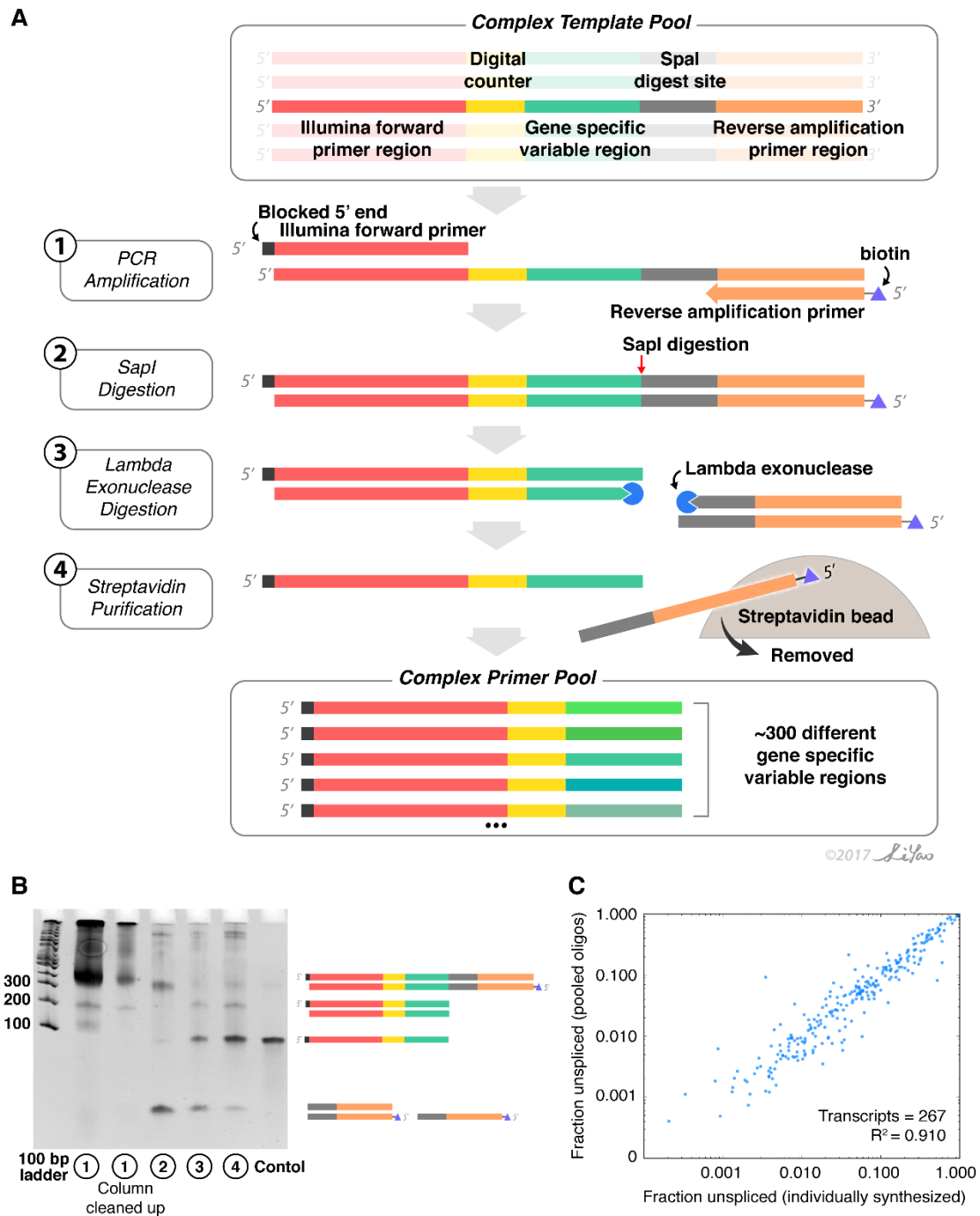


Figure 2.9: Complex pools of oligo mixtures can be generated to be used for priming in Splice-Seq. (A) Schematic showing the complex pool generation process. (B) 6% acrylamide gel showing the products of each purification step. 100 bp ladder used in for sizing. The PCR product is shown in 1. Column purified PCR products to its right.



SapI cleaved product shown in 2. Lambda exonuclease digested, single stranded products shown in 3. Streptavidin purified product shown in 4. Control lane shows pure mixture of oligo primers ordered individually. (C) Correlation between fraction unspliced generated from Splice-Seq libraries primed with either complex oligo mixture or individually synthesized and pooled mixture.

*Splice-Seq can be scaled to query thousands of target locations*

Having demonstrated that pooled oligonucleotide synthesis was a viable approach for generating Splice-Seq primers, we sought to determine whether this approach would be similarly robust in systems with far greater splicing complexity. To assess this, we turned to the fission yeast, *Schizosaccharomyces pombe*, an organism that shares many genetic properties with its distantly related cousin *S. cerevisiae*, but in many ways more closely resembles mammalian splicing. The *S. pombe* genome encodes thousands of introns, including thousands of genes with multiple introns, and widespread examples of alternative and aberrant splicing have been demonstrated (Awan, Manfredo, and Pleiss 2013). Using the lessons we learned in budding yeast, we designed a pool of ~4000 oligonucleotides, each targeting an individual intron within the *S. pombe* genome. Following the procedure outlined in Figure 2.9A, we amplified the *S. pombe*-specific primers, and then used them to generate replicate Splice-Seq libraries from RNA derived from a wild type strain of *S. pombe*. The results of these experiments were in line with the previously described experiments, with some important differences. As seen in Figure 2.10A, while a smaller fraction of the reads reflected splicing informative primer extension reactions, that subset of reads nevertheless provided similar information about splicing status across thousands of splicing events in *S. pombe* (Figure 2.10B). By contrast, for reasons that are yet unclear but remain under investigation, a much higher percentage of the reads generated in this

experiment were either unmappable or reflected the cloning of unextended primers. Similar artifacts plagued early attempts at Splice-Seq in *S. cerevisiae* but were rectified with changes to library purification steps; we expect similar iterations to yield increased purity for these *S. pombe* libraries as well.

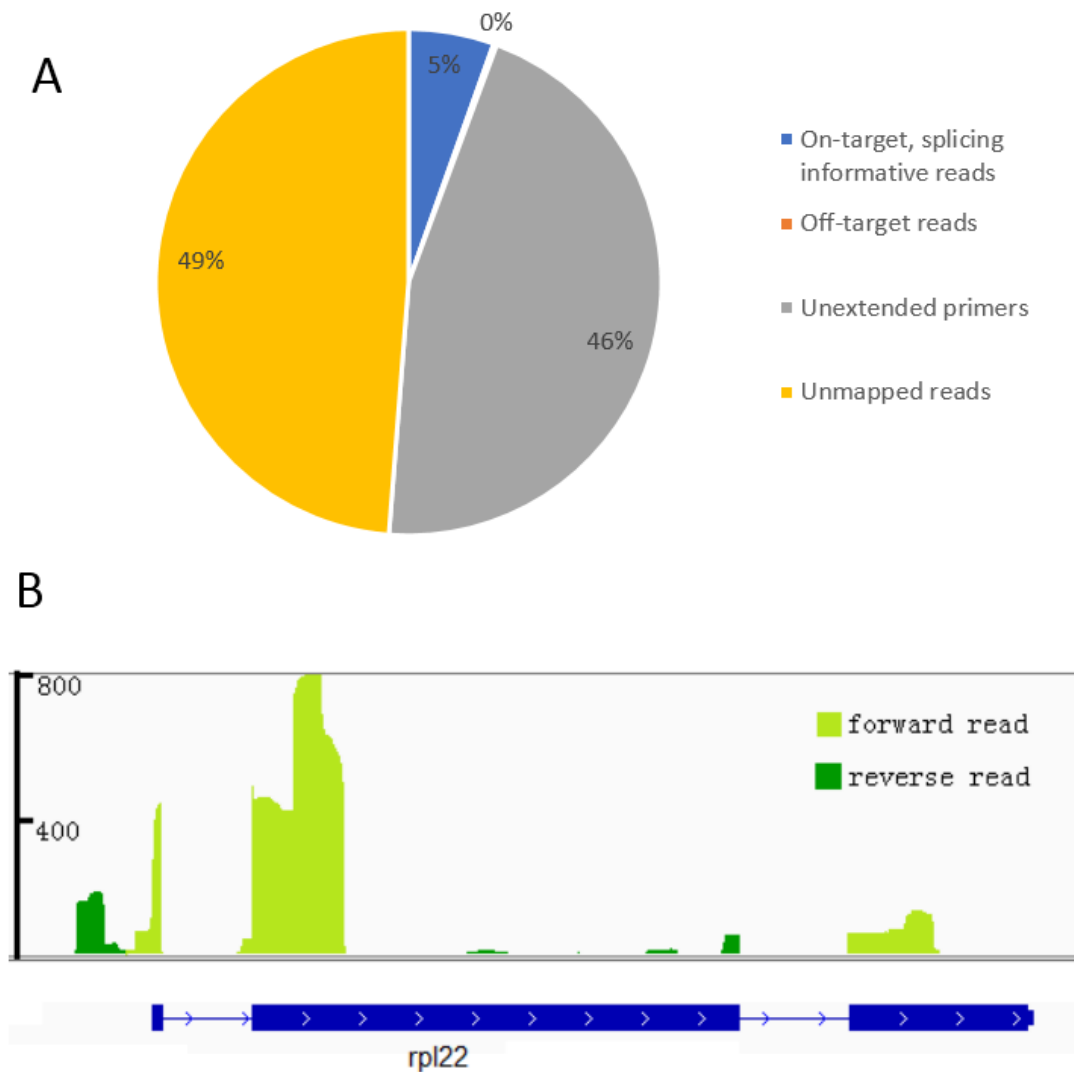


Figure 2.10: Splice-Seq can be upscale to query thousands of target locations. (A) mapping distributions for Splice-Seq in *S. Pombe*. (B) Example multi-intron gene coverage for rpl22.

## 2.4 Discussion:

Here we have described the development and implementation of a targeted sequencing method, which we call Splice-Seq, that was designed to measure changes in pre-mRNA splicing state with increased precision but decreased cost relative to traditional RNA-Seq. Whereas the majority of reads derived from most RNA-Seq experiments map directly to exonic regions and therefore lack direct information about splicing state, Splice-Seq targets these reads to splicing junctions through the use of user-selected primers during reverse transcription. Our initial experiments in the budding yeast *S. cerevisiae* demonstrated an increase in sensitivity of over two orders of magnitude as compared with RNA-seq (Figure 2.4D). We simultaneously described a method that allows for robust generation of pools of single stranded primers of user-selected identity that readily enable the application of Splice-Seq to any set of splicing events within any organism (Figure 2.9). Together, these tools should enhance the ability of researchers to interrogate a variety of aspects of pre-mRNA splicing.

An important example of the novel type of information that can be derived from the increased resolution provided by Splice-Seq is our demonstration of the ability to distinguish between fully unspliced and lariat intermediate versions of pre-mRNA. To our knowledge, no previous method has allowed for the *in vivo* differentiation of these species on a global scale. The ability to distinguish between these products *in vivo*, particularly when coupled to methods for rapid metabolic labeling of nascent RNAs (Rädle et al. 2013, Schwarzl et al. 2015, Oesterreich et al. 2016), provides a profound opportunity to kinetic behavior of the global complement of splicing substrates not only under standard conditions but also in the context of changes environmental or genetic conditions. Indeed, whereas the spliceosomal helicase Prp16 has long been associated with a structural rearrangement of the spliceosome necessary for the second chemical step, and while proteomic studies have only detected its stable presence in complexes

with the spliceosome after completion of the first chemical step, our demonstration here that strains harboring the *prp16-302* allele accumulate products consistent with both a first and second step defect suggest that the role of this factor may be more expansive than originally thought (Figure 2.7). Additional experiments will be necessary to determine whether the results presented here reflect a primary role for Prp16 during the first chemical step of splicing, or whether they reflect downstream consequences of this mutation. We expect the tools presented here will provide the opportunity to address both this question and others like it.

## 2.5 Materials and methods:

### *Strains and cell growth:*

The wildtype strain used in the experiments presented here was BY4741. The mutant *prp2-1* and *prp16-302* strains were both obtained from the Guthrie lab. BY4741 was paired with both *prp2-1-ts* and *prp16-302-ts* and underwent either heat or cold shift respectively before collection. Heat shifted strains were grown in YPD at 25 °C until saturation, then back diluted to OD 0.05 and grown until ~OD 0.7 at which point the temperature of the growth medium was raised to 37 °C by addition of an equal volume of fresh 50 °C YPD for 15 minutes before cells were collected by filtration. Similarly, cold shifted strains were grown in YPD at 30 °C until saturation, then back diluted to OD 0.05 and grown until ~OD 0.7 at which point the temperature of the growth medium was reduced to 16 °C by addition of an equal volume of fresh 0 °C YPD for 15 minutes before cells were collected by filtration. Cell pellets were frozen in liquid nitrogen and stored at -80 °C until RNA extraction.

### *Splice-Seq library preparation:*

RNA extraction: RNA was isolated and then DNase treated using standard methods according to Inada and Pleiss 2010.

cDNA synthesis and cleanup: 10 µg of DNase treated total RNA was used as input per library where 1 µg of gene specific primer pool (each gene specific primer at equimolar ratio) was added and annealed by heating at 70 °C for 1 minute, 65 °C for 5 minutes, then cooled down and held at 47 °C. An equivalent volume of MMLV reverse transcriptase enzyme mix containing 1 mM dATP, 1 mM dGTP, 1 mM dCTP, 0.4 mM aminoallyl-dUTP and 0.6 mM dTTP and appropriate buffers and salts was pre-heated to 47 °C and added to the primer-annealed RNA mix. Reactions were incubated at 47 °C for 3 hours, followed by heat inactivation at 85 °C for 5 minutes. Remaining RNA was

hydrolyzed by addition of ½ volume of 0.3M NaOH with 0.03M EDTA and incubated at 65 °C for 15 minutes. After neutralization with ½ (original) volume of 0.3M HCl, the cDNA was purified with a Zymo 5 column using 7X volume binding buffer (2M guanidinium-HCl, 75% isopropanol, Zymo Research C1003-50) following manufacturer's protocol. Purified cDNA samples were dried in a SpeedVac for further processing.

NHS ester biotin coupling: Purified cDNA pellets were resuspended in 18 µL of fresh 0.1M Sodium Bicarbonate pH 9, to which was added 2 µL of 0.1 mg/µL NHS-biotin (ThermoFisher 20217, dissolved in DMSO); reactions were incubated at 65 °C for 1 hour. cDNA was purified from unreacted NHS-biotin using Zymo columns, again using 7X volume binding buffer and following the manufacturer's protocol.

Streptavidin-biotin purification: Biotin-labeled cDNAs were purified in a reaction containing 20 µL of Dynabeads MyOne Streptavidin C1 (ThermoFisher 65601). Dynabeads were pre-washed twice in 500 µL of 1X binding and washing buffer (5 mM TrisHCl pH 7.5, 0.5 mM EDTA and 1M NaCl, as per manufacture's protocol) prior to use. Washed beads were resuspended in 2X binding and washing buffer and were combined with the purified cDNAs in a total of 100µL and allowed to rotate for 30 minutes at room temperature. Bound material was washed twice with 500 µL of 1X binding and washing buffer, then two more times with 100 µL of 1X SSC. To ensure purification of only single-stranded cDNAs, pellets were then incubated with 0.1M NaOH for two consecutive 10 minute washes. Finally, the bound material was washed 3 times with 1X TE before the purified cDNA was eluted by heating the sample to 90 °C for 2 minutes in the presence of 85 µL of 95% formamide, 10 mM EDTA. The eluate from this step was then purified again using a Zymo column using 7X volume binding buffer following the manufacturer's protocol.

First strand extension: The dN9-anchored Illumina reverse Nextera primers were

annealed to the first strand by heating the purified cDNA in the presence of 2  $\mu$ M primers to 65 °C for 5 minutes, then cooling down to room temperature. Klenow exo-fragment (NEB M0212S) was added and the reactions incubated for 5 minutes at room temperature, after which they were moved to 37 °C for 1 hour. Purification of the products of this reaction by streptavidin was accomplished according to the same procedure described above.

PCR amplification: Amplification of the reaction products was accomplished by using ¼ of the material generated as a template in a PCR reaction. Illumina Nextera (i5) and (i7) adapters were used in a standard 50  $\mu$ L PCR reaction with Phusion polymerase (ThermoFisher F530S). Cycling conditions were as follows: denaturation at 95 °C for 10 sec; annealing at 62 °C for 20 sec; and extension at 72 °C for 30 sec. Libraries typically required between 14 and 20 cycles of amplification, depending upon the amount of starting material. Amplified material was purified using 6% native PAGE stained with SyBr gold (ThermoFisher S11494), where material in the range from 200bp to 800bp was extracted.

*Complex oligo mix amplification method:*

PCR amplification: Using Phusion polymerase under standard PCR conditions, double-stranded amplicons were generated from the single-stranded oligos obtained from LCSciences using 14 rounds of PCR in a 400  $\mu$ L reaction containing: 1% of the pooled oligonucleotides from LCSciences as a template, an Illumina forward primer containing a C3 spacer (Integrated DNA Technologies) at its 5' end, and a reverse amplification primer containing a biotin-label at its 5' end. Cycling conditions were as follows: denaturation at 95 °C for 10 sec; annealing at 60 °C for 20 sec; and extension at 72 °C for 30 sec. Upon completion of this initial reaction, the entire reaction was used as a template to seed a larger (4mL) PCR reaction. For efficient amplification, this large

reaction was performed in two 96-well plates with 50  $\mu$ L in each well. Reaction conditions were identical to those described for the first reaction, and a total of 14 cycles were performed for this second amplification. Reactions were purified and concentrated by isopropanol precipitation under standard conditions.

Enzymatic digestions: The double-stranded amplicons were digested using SapI (NEB R0569L) according to manufacturer's protocol in a 150  $\mu$ L reaction containing 30  $\mu$ L of enzyme and incubated at 37 °C overnight. The products of this reaction were concentrated by ethanol precipitation under standard conditions. The resuspended DNA was then digested with lambda nuclease (NEB M0262L) at 37 °C for 2 hours according to manufacturer's protocol. The products of this reaction were purified using Zymo columns with 7X volume binding buffer .

Streptavidin-biotin purification: Removal of the reverse strand material was accomplished using 50  $\mu$ L of Dynabeads MyOne Streptavidin C1 according to the manufacturer's protocol. Importantly, the unbound supernatant fraction was retained as it contains the desired products. The recovered material was precipitated and verified using 6% native PAGE stained with SyBr Gold.

#### *Alignment, mapping and counting:*

Reads were sequenced on the NextSeq platform by Cornell Sequencing Center using 60bp (forward, 5' end of cDNA) +15bp (reverse, 3' end of cDNA) paired-end chemistry. The adapter sequences were clipped off the reads, and PCR duplicates were removed from the dataset by filtering out non-unique reads with respect to all base calls. Reads were aligned to the yeast genome (Engel et al. 2014) using the STAR aligner (Dobin et al. 2013) using the following alignment parameters: {--alignEndsType EndToEnd --clip5pNbases 7 0 --alignMatesGapMax 400 --alignSplicedMateMapLmin 16 --outSAMattributes All --alignSJDBoverhangMin 1 --outSAMmultNmax 1 --



outSAMunmapped Within KeepPairs --outFilterMismatchNmax 3}. Using a custom script, alignments were divided between continuous and gapped alignments via the occurrence of an "N" in the cigar string. Alignments crossing an intron and alignments crossing a junction were then separately counted on a per target basis using a modified version of htseq-count (Anders, Pyl, and Huber 2015). The second read of the paired end reads was used to determine the 3' termini of the cDNAs, which we observed to often terminate near the TSS, predicted highly structured areas of the RNA, or in the case of a lariat-intermediate-derived cDNA, near the branchpoint-A of the intron. To make heatmaps and meta-gene plots that quantify the 3' termination of cDNAs around these features, we used the deepTools ComputeMatrix command (Ramírez et al. 2016) in conjunction with a BigWig coverage file of the 3' terminating bases and bedfiles containing TSS-regions and the branchpoint-to-3'ss regions as previously annotated (Davis et al. 2000). The heatmaps and metaplots depicting the cumulative distribution of 3' termini were made by processing the deepTools computeMatrix output in order to successively sum up the counts of 3' termini while moving across the gene feature and expressing it as a percent of the total.

*Primer sequences used in the experiment:*

Primers were all ordered individually from Integrated DNA Technologies with the exception of the pooled oligo mixture, in which case LC Sciences was used.

Sequence of the reverse transcription primers consists of the following, in order from 5' to 3' end. The 3' end of the i5 adapter (after the index) of the Illumina Nextera primer sequence (5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG 3'), followed by 8 random nucleotides used for monitoring PCR artifacts (5'NNNNNNNN 3'), followed by the gene specific annealing regions as output by OligoWiz (24 – 26mers).

Sequence of the complex pooled oligo mixture consists of the same sequences as above, except with the additional of a SapI digestion site (5'CCTCGAAGAGC 3'), and addition of a reverse amplification region (5'ATTACGGCTCCTCGCTGCAG 3').

Illumina forward primer used to amplify complex mixture of oligo pool: (Biotin 5' CTGCAGCGAGGAGCCGTAATGC 3').

Reverse primer used to amplify complex mixture of oligo pool: (5' 3C spacer CTGCAGCGAGGAGCCGTAATGC 3').

First strand extension primer: (5' GTCTCGTGGGCTCGGAGATGTGTAT AAGAGACAGNNNNNNNNN Hexanediol spacer 3')

## **Chapter 3: Deciphering a Role for the Chromatin Remodeling Factor Bdf1 in pre-mRNA splicing**

### **3.1: Introduction**

Splicing and transcription are not separate cellular events but are deeply interconnected to each other. A variety of experiments over the past several years demonstrate that splicing factors are recruited to the nascent RNA co-transcriptionally (Bentley DL 2005, Neugebauer KM 2002, Tardiff DF 2006), yet the mechanisms by which they are recruited and by which their recruitment facilitates splicing remain poorly understood. In order to gain an understanding of the mechanistic details of co-transcriptional splicing, we set out to identify specific factors that were important for functionally coupling these processes together.

Laura Albulescu, a previous graduate student in the Pleiss lab, began the work on this project and had developed a genome-wide screen in *S. cerevisiae* where splicing efficiency was evaluated in the background of ~5500 unique gene mutations, covering almost the entire yeast genome. The method Laura developed used quantitative PCR to measure expression changes in pre-mRNA levels in the background of the mutants: increased levels of pre-mRNA in a strain suggested a defect in splicing, and thus implied the corresponding gene plays a role in splicing (Albulescu et al. 2012). The screen not only identified an array of known splicing factors, but also identified many candidates previously known only for playing roles in chromatin remodeling, transcription, or 3' end processing. One such candidate was Bdf1, a bromodomain containing factor that is a component of the transcription factor TFIID and a member of the SWR-C chromatin remodeling complex (Durant and Pugh 2006). The Bdf1 protein functions at the interface between transcription and chromatin remodeling. As a part of the TFIID complex, Bdf1 functions to recruit RNA polymerase II to TATA-less promoters through its interaction with Taf7 (Durant and Pugh 2007, Martinez-Campa et al. 2004, Pamblanco et al. 2001). In chromatin remodeling, the yeast SWR-C complex is

responsible for exchanging histone H2A with histone H2A.Z at promoters genome-wide (Krogan et al. 2003, Kobor et al. 2004, Zhang, Roberts, and Cairns 2005). Prior to Laura's work, Bdf1 had not previously been implicated in pre-mRNA splicing, and indeed the mechanistic basis for the splicing defect observed in strains lacking Bdf1 is unknown.

Interestingly, Bdf1 has a close homolog Bdf2 which emerged after the whole genome duplication event in yeast (Wolfe 2015). Neither protein is essential on their own, but they are synthetically lethal when both are deleted. Though most duplicated copies of genes from the whole genome duplication event quickly became non-functional, the fact that both Bdf1 and Bdf2 remained is cause for speculation that they acquired diverged functions. When examining the sequence however, the two proteins are actually highly similar, with 35% identical and 67% similarity, with both retaining two bromodomains and a C-terminal acidic domain. Nevertheless, there are also regions of differences between the two proteins where stretches of amino acids are present in one protein but not the other and vice versa (Figure 3.1). As the proteins are highly similar, it is of no surprise that Bdf2 can interact with Taf7 as well (Matangkasombut and Buratowski 2003). There is also evidence to suggest that they can at least have a partial overlap in function as Bdf2 is upregulated upon Bdf1 loss and can bind some Bdf1 specific promoters (Durant and Pugh 2007).

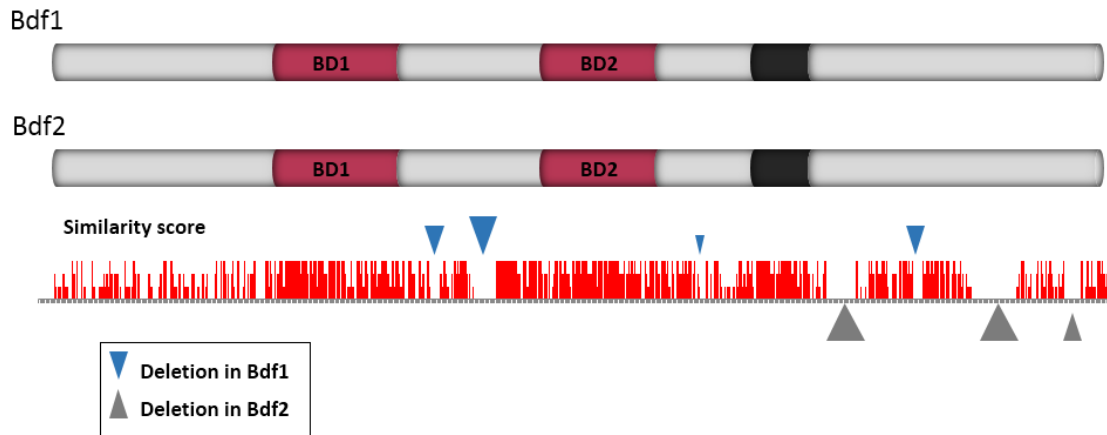


Figure 3.1: Bdf1 and Bdf2 are highly similar with stretches of amino acids that are different. Gene structures for Bdf1 and Bdf2 shown, with bromodomains highlighted in red and acidic domain in black.

Surprisingly, contrary to Bdf1, Bdf2 was not identified as a factor important for splicing in our previously mentioned genome-wide screen. Splicing sensitive microarrays performed on *bdf1Δ* confirm that deletion of Bdf1 leads to a global splicing defect while *bdf2Δ* shows normal splicing (Figure 3.2, Albulescu et al. 2012). This suggests that even though both proteins are implicated in transcription, only Bdf1 plays a role in splicing. This leads to the interesting question of how Bdf1 is involved in pre-mRNA splicing and what regions have diverged between the two proteins that enable this function.

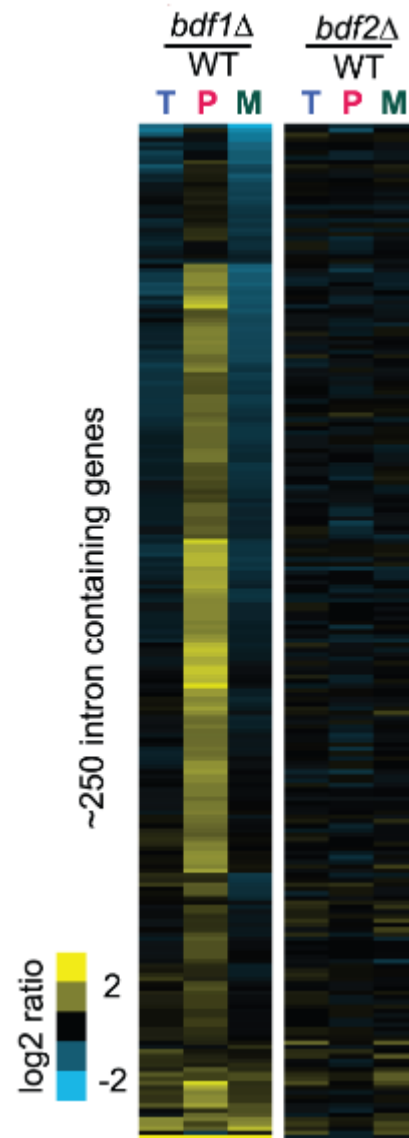


Figure 3.2: Bdf1 plays a role in splicing while Bdf2 does not. Splicing sensitive microarray performed on strains lacking Bdf1 show a marked increase in the pre-mRNA species whereas strains lacking Bdf2 largely remain unchanged. T P and M panels represent Total, pre-mRNA and mature mRNA. Adopted from Albulescu et al. 2012.

As an initial approach to examine this problem, the mechanism by which Bdf1 impacts splicing was examined. By monitoring U1 snRNP recruitment in the background of wildtype, *bdf1Δ* and *bdf2Δ* mutants through chromatin immunoprecipitation coupled to qPCR (ChIP-qPCR), it was observed that U1 snRNP recruitment is diminished upon Bdf1 deletion throughout different genomic regions of the ACT1 gene (Figure 3.3A), as well as throughout intronic regions of U3, RPL31B and UBC13 (Figure 3.3B). This suggests that the loss of Bdf1 is detrimental to the co-transcriptional recruitment of U1 snRNP. To have a global look at the genes impacted by Bdf1 loss and how it differs from Bdf2 loss, a more comprehensive ChIP-Seq experiment will be required.

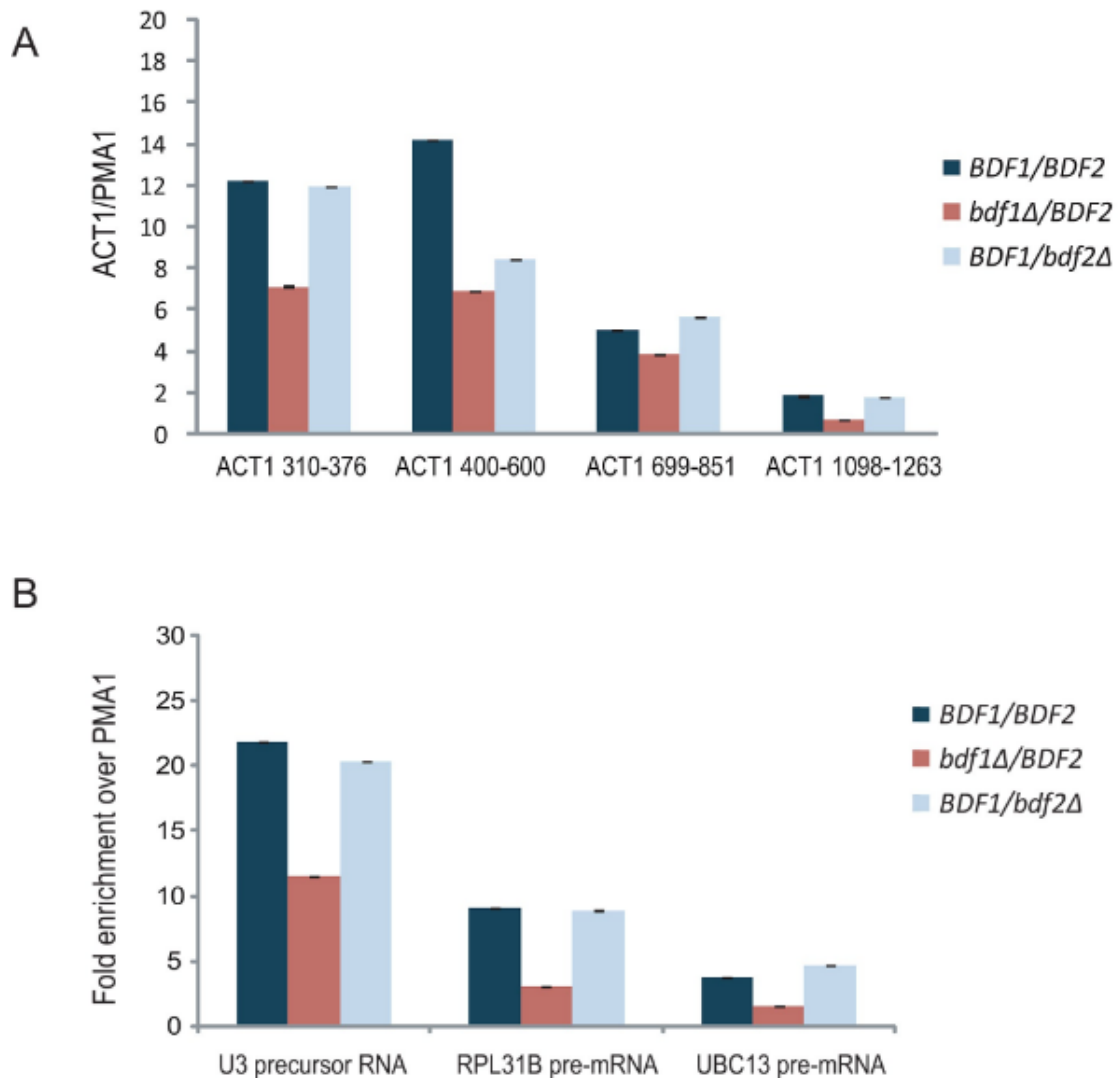


Figure 3.3: Loss of Bdf1 results in a defect in U1 snRNP recruitment as shown by ChIP-qPCR. Panel A shows *bdf1Δ* strain perform worse than both wildtype and *bdf2Δ* when it comes to measuring gDNA signal by qPCR from U1 ChIP throughout the actin gene. Panel B shows decreased signal from genomic precursor RNA locations on strains with Bdf1 loss as compared to wildtype and Bdf2 loss.

Given the high similarity between Bdf1 and Bdf2 but a lack of role in splicing for Bdf2, we hypothesize that Bdf1 and Bdf2 generally share redundant cellular functions in transcription, and that Bdf2 can compensate for the transcriptional function



of Bdf1 in a *bdf1Δ* strain. We further hypothesize that Bdf1 has an additional function where it also plays a critical role in spliceosomal component recruitment at the start of transcription that Bdf2 cannot substitute for. Given its role in transcription, the 5' biased nature of yeast introns (Figure 3.4), combined with the previously mentioned ChIP-qPCR results, we hypothesize that the spliceosomal component recruited by Bdf1 is the U1 snRNP.

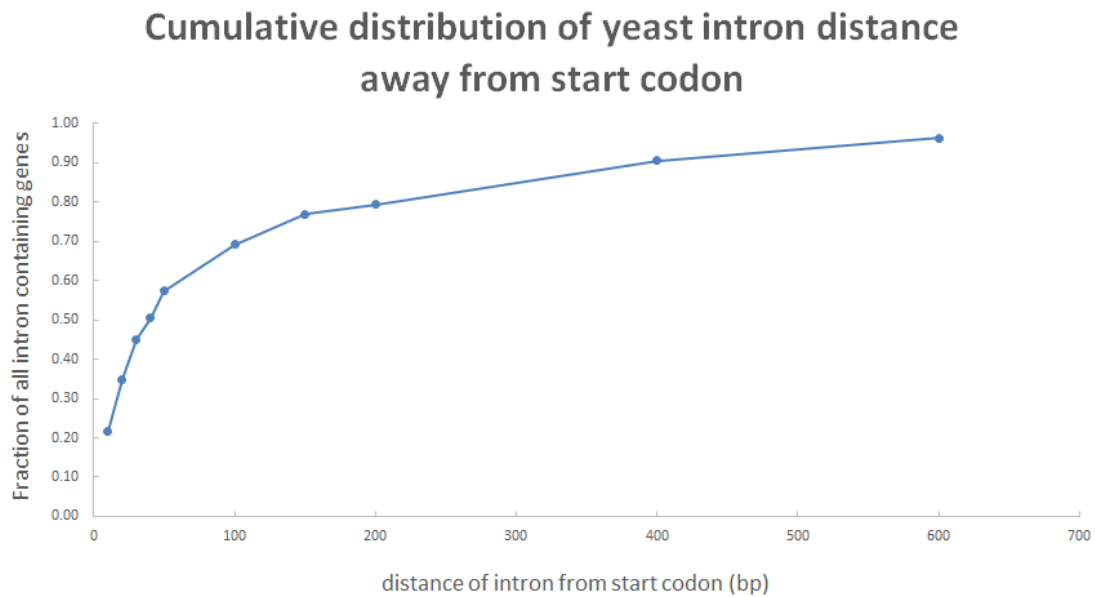


Figure 3.4: *S. cerevisiae* introns are 5' biased. Cumulative distribution plot showing the distance of the intron for intron containing genes in *S. cerevisiae* from the start codon to the intron location. 70% of all introns are within 100 basepairs from the start codon.

We present our hypothesis regarding Bdf1 and Bdf2 as the following model. Under normal wildtype conditions, Bdf1 is involved in chromatin remodeling, transcription initiation, as well as U1 snRNP recruitment (Figure 3.5, top panel). Under conditions where Bdf1 is deleted, Bdf2 is regulated and compensates for Bdf1's chromatin remodeling and transcription initiation roles. However, Bdf2 lacks the capacity to recruit U1 snRNP, thus resulting in a global splicing defect in *bdf1Δ* strains

(Figure 3.5, bottom panel). This chapter seeks to both identify the genome-wide consequences of a loss of Bdf1 on the recruitment of spliceosomal component U1 snRNP, as well as to identify regions of Bdf1 that are necessary for its interaction with U1 snRNP.

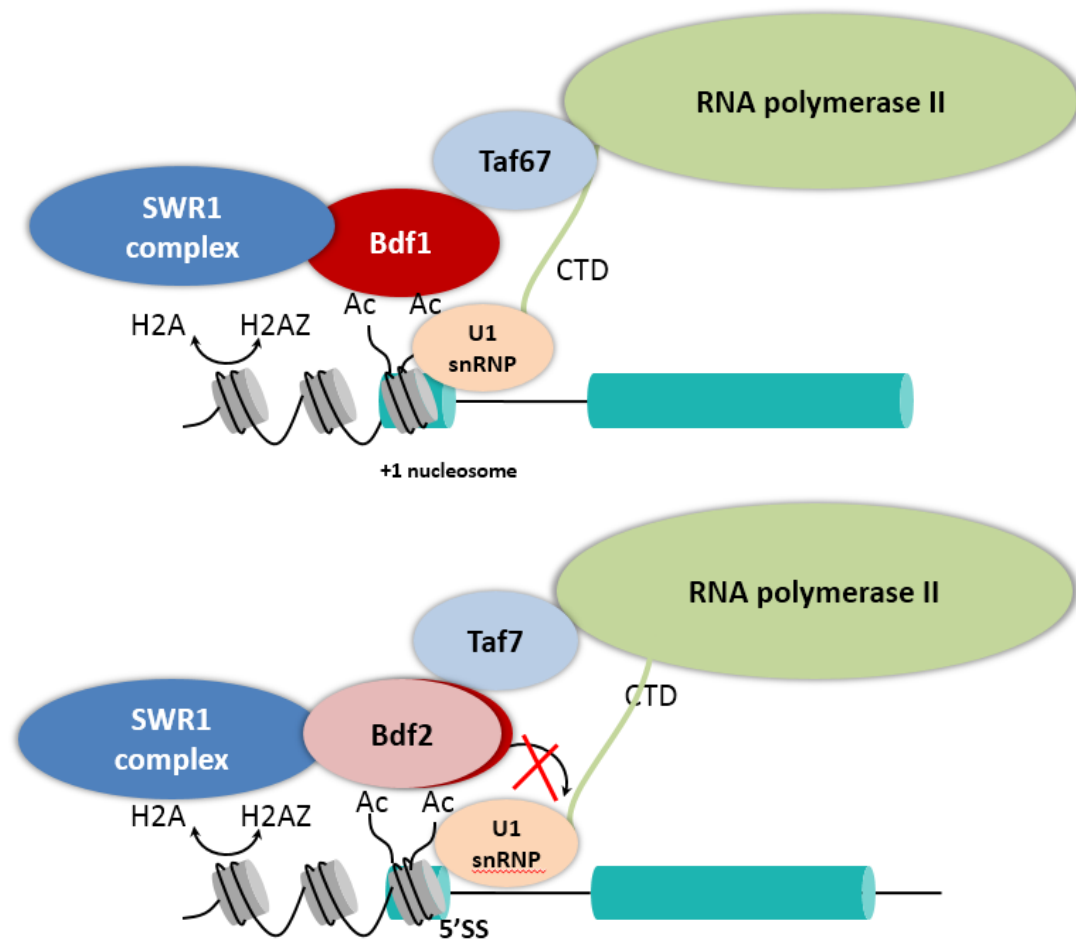


Figure 3.5: Proposed model for Bdf1's role in spliceosomal recruitment. Top panel: in the presence of Bdf1, U1 snRNP can be recruited co-transcriptionally by Bdf1. Bottom panel: Bdf2 is able to compensate the transcriptional role for Bdf1 but unable to recruit U1 snRNP, resulting in a global splicing defect.

### 3.2: Materials and methods

#### Strains:

Wildtype strain used is BY4741.

*bdf1Δ* and *bdf2Δ* come from Open Biosystems.

Mutant strains *bdf1\** generated by Laura Albulescu, as described by Albulescu LO 2012, Ph.D. dissertation.

#### Chip-Seq:

Cell lysis: Started with 300 mL total cells at OD 0.6 per sample. Pellet into a screw cap tube and add 500 uL of cold glass beads 0.5mm in diameter to each tube. Add 1 ml lysis buffer (50 mM Hepes (pH 7.5), 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% NaDeoxycholate, 1 Protease Inhibitor Tablet from Roche) and bead-beat for 15\*2 minutes with 1 minute intervals (for a total of 44 minutes). Puncture the bottom of the screw cap tube using a hot 21-gauge needle. Place tube on top of a 5ml syringe barrel inside a 15 ml Falcon tube. Centrifuge at 1000\*g at 4 °C for 1 minute to recover the lysed cells. Transfer samples to a new 1.7ml Eppendorf tube. Spin for 15 min at 14K rpm and 4 °C and discard supernatant. Resuspend samples in 1 mL of Lysis buffer. Transfer to 15 mL conical tubes.

Sonication: Using the Bioruptor sonicator, sonicate at high settings for 20 seconds with 1 minute off intervals for 25 cycles total. Ensure samples are in ice bath as heat can build up from the sonication. Sonication setting can vary, and shearing can be easily checked by test samples and gel electrophoresis. Transfer samples into clean 1.7 mL eppendorf tubes in 100 uL fractions. Spin eppendorf tube for 15 min at 14K rpm at 4 °C. Transfer supernatant to a new pre-chilled 1.7 mL eppendorf tube and spin for another 15 minutes at 14K rpm at 4 °C. Transfer supernatant to a new tube. Save a 10ul aliquot which will be your 1% Input. Add 90 ul of Elution buffer to the 1% Input and flash freeze. Store in the -20C.

Immunoprecipitation: Add Antibody to the samples and rotate at 4 °C for 2h. Prepare the resin (Protein A/G-agarose, cat #SC-2003 from Santa Cruz), 75 uL of slurry for 25 uL of resin for each tube. Washed 3 times with 1 mL Lysis buffer. After each wash spin down for 30 sec at 14k rpm in a tabletop microcentrifuge and let settle on ice for 1 minute before pipetting the wash buffer off. Keep on ice until use. Right before use, resuspend gently by pipetting up and down in lysis buffer of 95 uL per sample. Add 95 uL slurry to each tube. Incubate for 2 more hours at 4 °C then collect beads by spinning for 1 min at 2000 rpm. Remove the rest of the supernatant and wash the beads for 15 minutes at 4 °C repeated with the following buffers in sequence. 2 washes with 1 ml Lysis buffer (without protease inhibitors), 2 washes with 1 mL washing buffer (10 mM TrisHCl (pH 8), 250 mM LiCl, 0.5% NP-40, 0.5% NaDeoxycholate, 1 mM EDTA) +360 mM NaCl, 2 washes with 1 mL washing buffer, 2 washes with 1 mL 1X TE buffer. Elute in 100 uL elution buffer (50 mM Tris HCl (pH 8), 5 mM EDTA, 1% SDS, 25 mM EGTA) and incubate resin for 30 minutes at 65 °C with constant tapping every 5 minutes. Spin for 5 sec at 14K rpm and transfer supernatant to a new eppendorf tube. This will be your IP fraction. Incubate IP and 1% Input tubes overnight at 65 °C to reverse crosslinking.

Recovery of DNA fragments: Incubate the overnight samples with 12.5 uL of 20 mg/ml proteinase K for 2h at 42 °C. Use Cycle Pure kit from Omega Bio-Tek by following their instructions using the equilibration buffers. Elute in 40 uL. Quantify by Qubit using HS dsDNA kit. Standard library prep can then be performed, consisting of end repair, A-tailing, adapter ligation, library PCR amplification, size selection and sequencing.

#### *Genetics screen:*

Mutant library extraction and cDNA synthesis were done robotically according

to Albulescu et al. 2012. Genetics screen by sequencing was done by following the protocol from Larson et al. 2016. Data analysis for sequencing starts with reads demultiplexed using a combination of Nextera-specific indices and custom plate-specific barcodes (Mamanova et al. 2010) within the insert read. The bwa-mem aligner (Li and Durbin 2009) was then used to align reads to a custom index containing both the spliced and unspliced isoforms of the target U3 gene. Alignments were then reported as counts with respect to plate and well numbers for the *bdf1*\* mutants.

### 3.3: Results and Discussion

#### 3.3.1: ChIP-Seq of Bdf1 mutants

Preliminary ChIP-qPCR data suggested that U1 snRNP recruitment was diminished upon Bdf1 deletion throughout different genomic regions of the ACT1 gene, as well as throughout intronic regions of U3, RPL31B and UBC13 (Albulescu et al. 2012). In order to assess whether this is a genome-wide phenomenon, we performed ChIP-Seq on wildtype, *bdf1Δ* and *bdf2Δ* strains where the Yhc1 (U1C) gene of U1 snRNP was Tap-tagged to test for co-transcriptional U1 occupancy (see methods). As a control for data quality, we also immunoprecipitated against Rpb3, an RNA polymerase II subunit. We found that on average, in all 3 strains, U1 was immunoprecipitated and peaks at ~500bp downstream of the transcription start site (TSS) for intron containing genes. What is unclear is how the 3 different strains differ in degree in their ability in U1 snRNP recruitment. According to previously presented ChIP-qPCR data, wildtype and *bdf2Δ* strains should be equal in their ability to recruit U1 snRNP while *bdf1Δ* strains would perform significantly worse. Based on the data we obtained, *bdf1Δ* is apparently worse at U1 snRNP recruitment than wildtype, which is consistent with previously mentioned ChIP-qPCR result. However, *bdf2Δ* appears to have an even stronger defect in recruiting U1 snRNP, which contradicts with previously presented data that its deletion did not result in a splicing defect, and that it should have no effect on U1 snRNP recruitment (Figure 3.5).

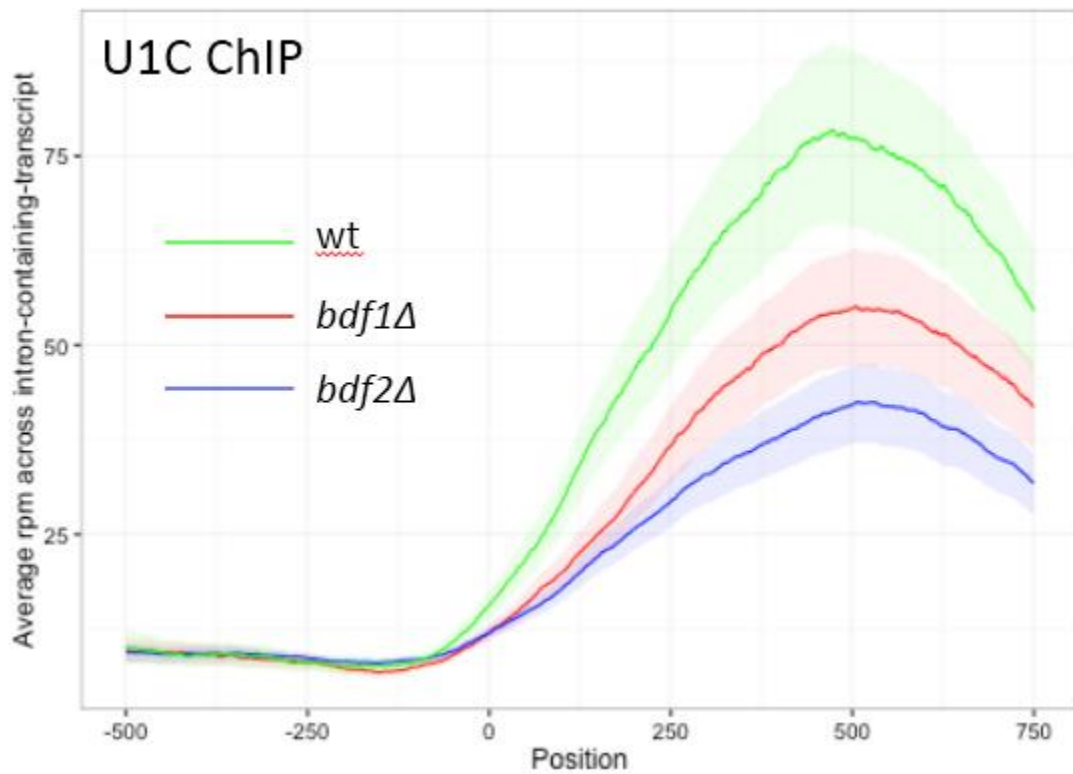


Figure 3.5: ChIP-Seq of U1C is apparently inconsistent with previously published ChIP-qPCR data. U1C ChIP-Seq shows less U1C pull down from *bdf1Δ* than wildtype, but more than that of *bdf2Δ*. Position 0 is aligned to the transcription start site.

To examine this discrepancy with previously published data, we look to eliminate the possibility that the conflicting answer we obtained is due to poor normalization. Upon observation of the ChIP-Seq profile for Rpb3, we see that RNA polymerase II pull down is not equally distributed in signal intensity even after correcting for sequencing depth. In the *bdf2Δ* strain, an apparent higher signal among gene bodies suggested more Rpb3 was pulled down than in the *bdf1Δ* strain, both of which have lower signal than wildtype (Figure 3.6). Taken at face value, if we assume Rpb3 pull down data can be used to normalize for the amount of pull down in U1 snRNP, we actually end up with an even lowered U1 snRNP pull down profile in *bdf2Δ*

as compared with either wildtype or *bdf1Δ*, making the observed discrepancy gap between ChIP-Seq and ChIP-qPCR wider, and not narrower.

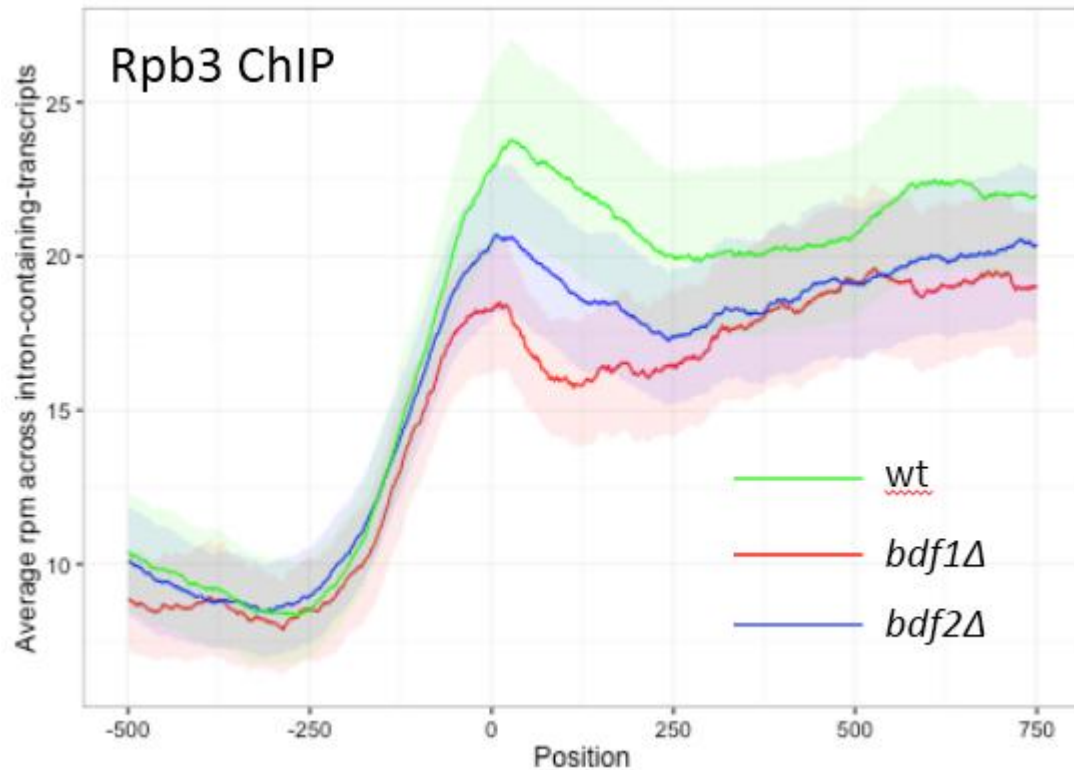


Figure 3.6: RNA Polymerase II ChIP-Seq data show the three strains, wildtype, *bdf1Δ* and *bdf2Δ*, pull down Rpb3 at unequal levels throughout intron containing genes.

Position 0 is aligned to the transcription start site.

Due to the noisy nature of antibody binding, normalization for the comparison of data between different ChIP-Seq experiments is difficult. It is widely known that uniform genome-wide occupancy increases or decreases can be missed (Bonhoure et al. 2014). Relative peak height, even though corrected to read depth, can still be misleading as there is no way to control for differences in antibody binding due to the different mutants used in different experiments.

In order to overcome this difficulty, a better normalization method is required. Such a method would need be able to perform cross sample comparisons of occupancy



levels for a set of loci of interest. Unless normalization methods that attempt to fix experimental variation after sequencing, this method would be performed before precipitation so that experimental noise is carried throughout the process. One such method, spike adjustment procedure (SAP), consists of an additional step prior to immunoprecipitation where a constant, low amount of a single batch of chromatin from a foreign organism is added to all experimental samples as an internal control (Bonhoure et al. 2014). This method will allow the adjustment of each experiment to the internal control and thus making their occupancy directly comparable to each other.

As a logical next step of the experiment, we have already obtained a *S. pombe* Chp1-Tap tagged strain (Verdel 2004 et al.) to be used for normalization. We should then repeat the previously described ChIP-Seq experiments with small amounts of *S. pombe* chromatin present, normalize all datasets to the internal *S. pombe* chromatin control, then be able to observe whether *bdf1Δ* does fail globally in recruiting U1 snRNP.

### **3.3.2 Forward genetics screen of Bdf1 mutants**

In order to decipher which part of the Bdf1 protein interacts with U1 snRNP, we made use of a library of ~1500 mutant strains of *bdf1\**, where \* denotes a different mutant in Bdf1, and screened the library for mutants which behaved like *bdf1Δ*. The library was prepared by Laura Albulescu during her work as a graduate student in the Pleiss lab (Albulescu LO, 2013 Ph.D. dissertation). Summarized briefly, the open reading frame of Bdf1 was mutagenized using Mutazyme II technology, where an average of 3-9 single nucleotide mutations had been introduced into each mutant ORF, those were then placed into a plasmid surrounded by the native Bdf1 genomic UTRs. This plasmid was then placed into a strain where genomic *BDF1* and *BDF2* were deleted, and a URA marked plasmid copy of Bdf2 was shuffled out on 5-FOA. This

method ensured that the only functional Bdf copy in the cell would be from *bdf1\**. The mutants of this library were individually arrayed in 384 well plates for easier subsequent manipulation with robotics.

To test the library of *bdf1\** for mutants that behaved like *bdf1Δ*, we employed a sequencing based genetics screen developed by Amy Larson and Benjamin Fair (Larson et al. 2016). Briefly summarized, U3 precursor mRNA levels were used as a proxy for splicing efficiency, where an increased precursor level indicated a failure in the splicing pathway. With the aid of robotics, the ~1500 *bdf1\** mutant library were grown in YPD under normal conditions, then had RNA extracted and reverse transcription performed to synthesize cDNAs using random primers. Inward facing primers containing sequencing amplifiable overhangs annealed to regions in exon 1 and exon 2 of U3 respectively were designed and used to amplify cDNAs synthesized from both the mature mRNA as well as the pre-mRNA. Custom barcoded PCR amplifications were then performed so each mutant would have a uniquely identifiable barcode. The sequencing library was then pooled and sequenced by Illumina Nextseq (Figure 3.7, modified from Larson et al. 2016).

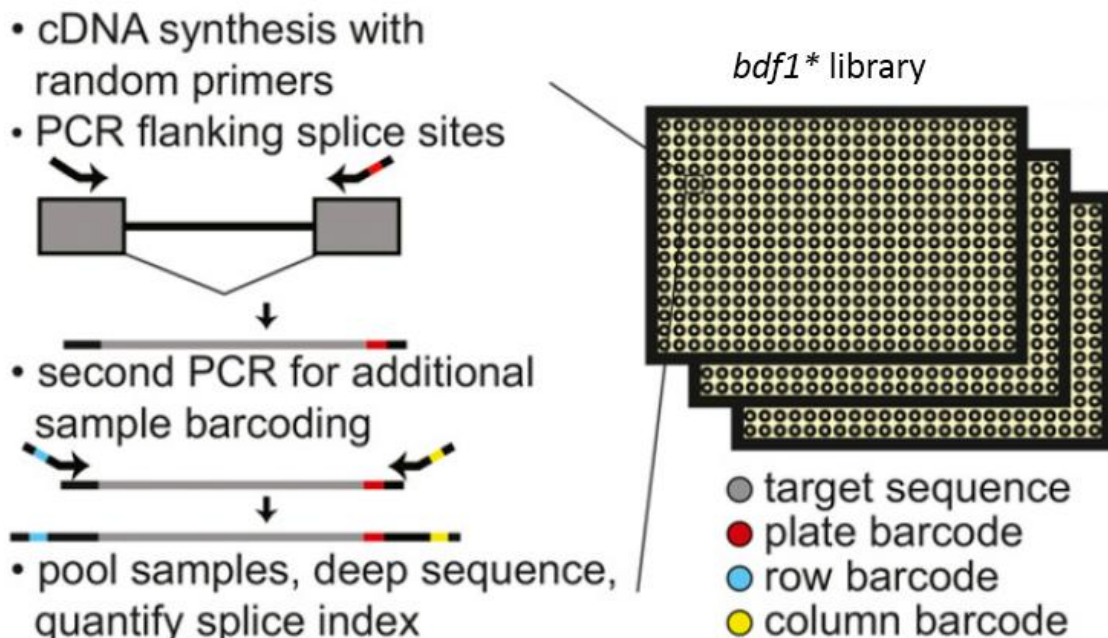


Figure 3.7: Schematic showing the forward genetics screen for *bdf1*\*. The mutants are arrayed on multiple 384 well plates, and RNA can be robotically extracted and reverse transcribed into cDNA. Multiple rounds of PCRs are then performed to append barcodes and the library is then pooled and sequenced. Modified from Larson et al.

2016

The reads from sequencing are splicing informative because they will either read into the intron-exon boundary, indicating the presence of a pre-mRNA, or read into the exon-exon boundary, indicating the mature species. We can use this information to calculate splice indices for each mutant and find *bdf1*\* mutants that have a high fraction of precursor U3 mRNA, indicative of a defect in splicing.

Previously published qPCR results indicate that U3 has a splice efficiency of ~97.5% (Pleiss et al. 2007), and so we designed a sequencing experiment for which each mutant strain was sampled with ~10,000 total reads, a level of sequencing high enough that even the unspliced species should be sampled at high enough read depth to provide reproducible values. To our surprise and disappointment, when considering the combined data from the ~1500 mutant strains of this experiment, a total of ~15 million reads were obtained that map to the mature species of U3 mRNA, whereas fewer than ~1500 reads were obtained that map to the precursor U3 mRNA. This paucity of unspliced reads was not only inconsistent with our expectations for the level of U3 pre-mRNA that would be present in the cell, but was also so low as to preclude any meaningful calculations of changes in relative splicing efficiency for any of the mutant samples.

We hypothesize the discrepancy in pre-mRNA detection comes from the inability of reverse transcriptase to make a full-length cDNA copy of the U3 pre-mRNA, possibly due to intronic structures that would preclude reverse transcriptase activity.

Whereas in qPCR, U3 pre-mRNA levels were measured by using primer pairs that reside at the 3' end of the intron and beginning of exon 2 respectively, with only a few basepairs between them, the primers in the previously designed sequencing experiment would require the entire length of the intron present to be able to amplify the full pre-mRNA sequence.

To test this hypothesis, saturating PCR was performed on the *bdf1*\* strain using the primer pair from the screen and analyzed by gel electrophoresis. The expected product size for the mature mRNA is 163 basepairs and the expected product size for the pre-mRNA is 320 basepairs. As is evident from the gel image, no pre-mRNA product is detectable, in spite the robust detection of the mature mRNA, and in line with our hypothesis (Figure 3.8).

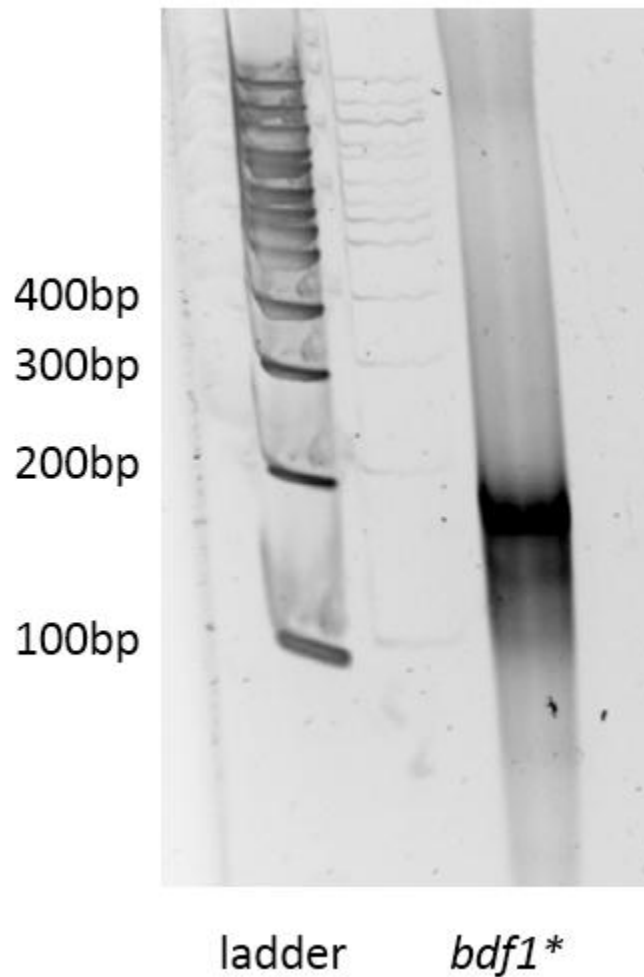


Figure 3.8: No expected product size observed for the pre-mRNA of *bdf1\** mutants using exon flanking primers. 6% PAGE showing the product of saturating PCR on *bdf1\** mutants next to 100 basepair ladder. Expected size of mature mRNA is 163 basepairs and pre-mRNA is 320 basepairs. Only the mature mRNA is observed in *bdf1\** strains.

To circumvent this problem, we plan to modify the screen protocol in a way that does not require the full-length cDNA to be made for it to be amplifiable. Instead of amplifying the cDNA product directly, we will incorporate the first strand extension strategy employed in chapter 2 of this dissertation, namely using an Illumina Nextera

sequence containing, randomer primed, 3' end blocked oligo in conjunction with Klenow to append the Illumina Nextera sequence onto the end of the newly synthesized cDNA. This will allow the amplification of any length cDNA products produced with a combination of the exon 2 U3 primer and a forward Illumina primer, even those that do not contain the full-length pre-mRNA. To prevent amplification of excess extension primers, blocked sequences complementary to the extension primer can be added in excess after the extension reaction, which will effectively remove the pool from being amplifiable, a strategy employed in the QuantSeq library preparation method (Moll et al. 2014).

Repeating the screen with the above-mentioned modifications should allow us to gain useful information on *bdf1*\* mutant phenotypes, and identify those mutants that phenocopy *bdf1Δ* strains. Once identified, we can identify the locations of the mutations within those mutant strains by sanger sequencing, and therefore find out which part of the protein is necessary for U1 recruitment.

## Future Directions

### Splice-Seq:

The development of the Splice-Seq technique should be further applied to higher systems with more complex intronomes. A natural next step is to attempt this in *S. pombe*, which has ~5000 introns compared with the ~300 in *S. cerevisiae*. A complex pool of primers targeting those ~5000 introns were ordered and amplified as described in Chapter 2 for *S. pombe* and Splice-Seq libraries were made using those primers. The overall mapping data for the preliminary Splice-Seq run in *S. pombe* is shown in figure S1.

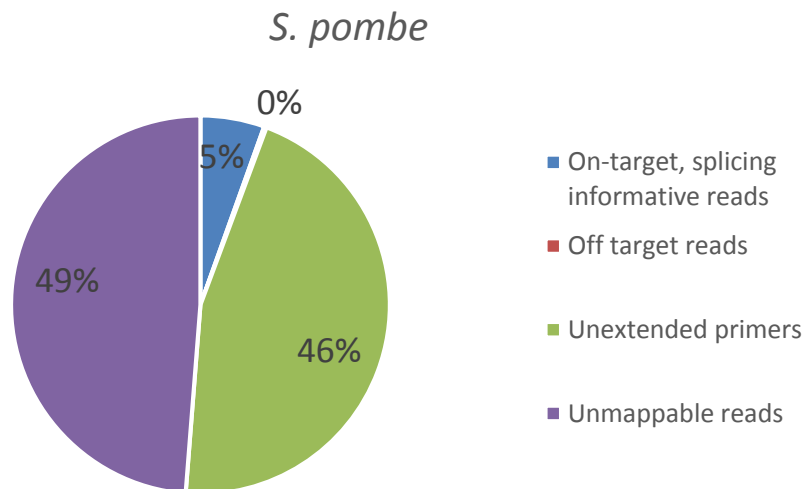


Figure S1: Preliminary mapping statistics of Splice-Seq in *S. pombe*.

As is evident from Figure S1, the overall read mapping distribution appears to have a larger fraction of both unmappable reads as well as unextended primers when compared to the mapping statistics of *S. cerevisiae*, shown in Figure 2.1B. Importantly, despite the increase in the number of unextended primers and off-target reads, there is still 5% of reads that are on-target and splicing informative, which is what we would have expected them to be in the first place. One such example is shown for the multi-

intronic rpl22 gene in figure S2.

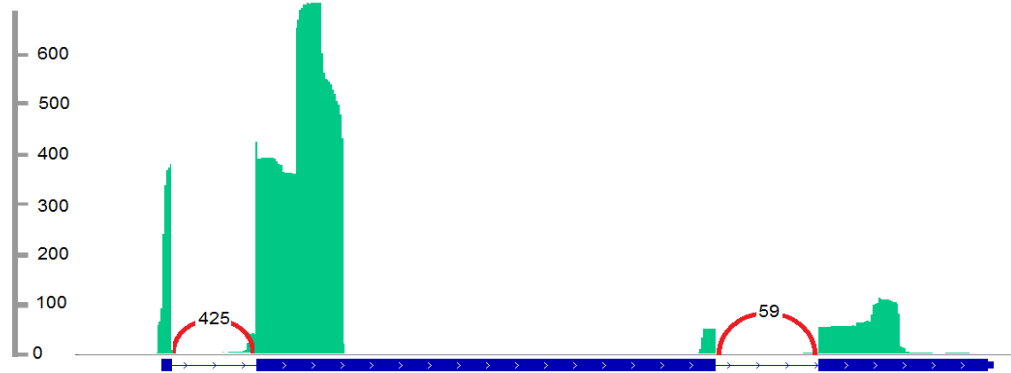


Figure S2: Splice-Seq mapping coverage for rpl22 in *S. pombe*. Green indicates mapped reads, red arc denotes mature reads (reads that map across exons).

I hypothesize that the increase in off-target reads when compared to *S. cerevisiae* is due to the increased complexity of the *S. pombe* intronome. To circumvent this problem, I propose several potential fixes. The first of which is to decrease mis-annealing events by raising the temperature and make use of a thermostable reverse transcriptase in the RT reaction instead of the MMLV enzyme used in the current method. Commercial thermostable reverse transcriptase enzymes have been previously tested in our lab (ThermoScript, Invitrogen 12236-014) to be able to synthesize cDNA at well beyond the average annealing temperatures of ~55 °C (up to 65 °C has been tested, unpublished data) for which the pool of gene specific reverse transcription primers were designed for. The increase in temperature should decrease the amount of mis-priming and decrease the fraction of reads that are off-target or unmapped due to having too many mismatches because of mis-priming. In order to combat the increased complexity of the reverse transcription reaction and make the substrates easier to find each other, the input RNA used for Splice-Seq in *S. pombe* can also be fragmented instead of using intact total RNA as in the case of the current method. Fragmenting the RNA will in theory allow the targets to flow better and find the targets easier. To



decrease the amount of background priming, it is also worth trying to use selected RNA as the input to Splice-Seq libraries instead of total RNA. Either poly-A selection or ribo-depletion will enable significantly reduced background levels at the cDNA synthesis stage.

In addition to optimizing Splice-Seq in higher systems, more work can be done in *S. cerevisiae* to further our understanding of the rates of splicing. Coupled with metabolic labeling where nascent RNA can be obtained through a time dependent manner, both the overall rate of splicing, as measured by the change in the amount of intronic reads and mature reads over time, as well as the rates of splicing for each of the two chemical steps of splicing, as measured by the change in the amount of totally unspliced pre-mRNA reads, lariat intermediate reads and mature mRNA reads over time, can be obtained by plotting that change over time. Preliminary work is already underway in this area, and we hope that by doping in a small amount of control to normalize the samples across different time points, we will soon be able to calculate absolute rates of splicing with the help of Splice-Seq.

#### *BDF1:*

To gain a further understanding of what is the unique role that Bdf1 plays in splicing in addition to its roles in transcription, I would like to identify a mutant in Bdf1 that phenocopies *bdf1Δ* strain. Repeating the screen to the library of mutants that Laura had prepared using the modifications mentioned in Chapter 3 should allow me to identify such a mutant, and subsequent identification of the mutation will allow insight into which region of the Bdf1 protein is important for such an interaction.

To assess how Bdf1 affects spliceosomal recruitment, further ChIP-Seq experiments will also be carried out. As described in Chapter 3, in order to perform a good ChIP-Seq experiment, proper normalizations with a foreign DNA sample input is

required and will allow the direct comparison between samples, and allow for assessment of different U1 binding profiles in mutants of *bdf1*.

## Credits

### Chapter 1:

Writing: Hansen Xu

Figures: Hansen Xu

### Chapter 2:

Writing: Hansen Xu and Jeffrey Pleiss

Ideas: Hansen Xu and Jeffrey Pleiss

Experiments: Hansen Xu

Data Analysis: Hansen Xu, Benjamin Fair (alignment), Zach Dwyer (read counting)

Figures: Hansen Xu and Li Yao

### Chapter 3:

Writing: Hansen Xu

Ideas: Hansen Xu, Jeffrey Pleiss, Laura Albulescu

Experiments: Hansen Xu (ChIP-Seq, Screen-Seq), Laura Albulescu (ChIP-PCR, generation of *bdf1*\* mutants).

Data Analysis: Hansen Xu

Figures: Hansen Xu, Benjamin Fair, Laura Albulescu

## Bibliography

1. Albulescu, L.-O. *et al.* A Quantitative, High-Throughput Reverse Genetic Screen Reveals Novel Connections between Pre-mRNA Splicing and 5' and 3' End Transcript Determinants. *PLoS Genetics* **8**, e1002530 (2012).
2. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
3. Anczuk ów, O. & Krainer, A. R. Splicing-factor alterations in cancers. *RNA* **22**, 1285–1301 (2016).
4. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
5. Änk ö, M.-L. Regulation of gene expression programmes by serine-arginine rich splicing factors. *Semin. Cell Dev. Biol.* **32**, 11–21 (2014).
6. Ares, M., Grate, L. & Pauling, M. H. A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* **5**, 1138–1139 (1999).
7. Awan, A. R., Manfredo, A. & Pleiss, J. A. Lariat sequencing in a unicellular yeast identifies regulated alternative splicing of exons that are evolutionarily conserved with humans. *PNAS* **110**, 12762–12767 (2013).
8. Baralle, F. E. & Giudice, J. Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology* **18**, 437–451 (2017).
9. Bentley, D. L. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Current Opinion in Cell Biology* **17**, 251–256 (2005).

10. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* **74**, 3171–3175 (1977).
11. Bitton, D. A. *et al.* LaSSO, a strategy for genome-wide mapping of intronic lariats and branch points using RNA-seq. *Genome Research* **24**, 1169–1179 (2014).
12. Bonhoure, N. *et al.* Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Research* **24**, 1157–1168 (2014).
13. Booth, G. T., Wang, I. X., Cheung, V. G. & Lis, J. T. Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast. *Genome Res* **26**, 799–811 (2016).
14. Bowne, S. J. *et al.* Identification of Disease-Causing Mutations in Autosomal Dominant Retinitis Pigmentosa (adRP) Using Next-Generation DNA Sequencing. *Investigative Ophthalmology & Visual Science* **52**, 494 (2011).
15. Carmel, I. Comparative analysis detects dependencies among the 5' splice-site positions. *RNA* **10**, 828–840 (2004).
16. Chanfreau, G. & Jacquier, A. An RNA conformational change between the two chemical steps of group II self-splicing. *EMBO J.* **15**, 3466–3476 (1996).
17. Chen, H.-C. & Cheng, S.-C. Functional roles of protein splicing factors. *Biosci. Rep.* **32**, 345–359 (2012).
18. Chen, M. & Manley, J. L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**, 741–754

(2009).

19. Chiou, N.-T. & Lynch, K. W. Mechanisms of spliceosomal assembly. *Methods Mol. Biol.* **1126**, 35–43 (2014).
20. Clark, T. A., Sugnet, C. W. & Ares, M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907–910 (2002).
21. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613–619 (2008).
22. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology* **17**, (2016).
23. Coombes, C. E. & Boeke, J. D. An evaluation of detection methods for large lariat RNAs. *RNA* **11**, 323–331 (2005).
24. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188–1190 (2004).
25. Crotti, L. B. & Horowitz, D. S. Exon sequences at the splice junctions affect splicing fidelity and alternative splicing. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18954–18959 (2009).
26. David, C. J., Boyne, A. R., Millhouse, S. R. & Manley, J. L. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes & Development* **25**, 972–983 (2011).
27. Davis, C. A., Grate, L., Spingola, M. & Ares, M. Test of intron predictions reveals novel splice sites, alternatively spliced mRNAs and new introns in

- meiotically regulated genes of yeast. *Nucleic Acids Res.* **28**, 1700–1706 (2000).
28. Deckert, J. *et al.* Protein composition and electron microscopy structure of affinity-purified human spliceosomal B complexes isolated under physiological conditions. *Mol. Cell. Biol.* **26**, 5528–5543 (2006).
  29. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
  30. Dreumont, N. & Séraphin, B. Rapid screening of yeast mutants with reporters identifies new splicing phenotypes. *FEBS J.* **280**, 2712–2726 (2013).
  31. Durant, M. & Pugh, B. F. Genome-wide relationships between TAF1 and histone acetyltransferases in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **26**, 2791–2802 (2006).
  32. Durant, M. & Pugh, B. F. NuA4-directed chromatin transactions throughout the *Saccharomyces cerevisiae* genome. *Mol. Cell. Biol.* **27**, 5327–5335 (2007).
  33. Engel, S. R. *et al.* The reference genome sequence of *Saccharomyces cerevisiae*: then and now. *G3 (Bethesda)* **4**, 389–398 (2014).
  34. Fabrizio, P. *et al.* The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Mol. Cell* **36**, 593–608 (2009).
  35. Faustino, N. A. & Cooper, T. A. Pre-mRNA splicing and human disease. *Genes & development* **17**, 419–437 (2003).
  36. Galej, W. P. *et al.* Cryo-EM structure of the spliceosome immediately after branching. *Nature* **537**, 197 (2016).
  37. Gencheva, M. *et al.* Nuclear retention of unspliced pre-mRNAs by mutant DHX16/hPRP2, a spliceosomal DEAH-box protein. *J. Biol. Chem.* **285**, 35624–

- 35632 (2010).
38. Gould, G. M. *et al.* Identification of new branch points and unconventional introns in *Saccharomyces cerevisiae*. *RNA* **22**, 1522–1534 (2016).
  39. Hang, J., Wan, R., Yan, C. & Shi, Y. Structural basis of pre-mRNA splicing. *Science* **349**, 1191–1198 (2015).
  40. Herold, N. *et al.* Conservation of the protein composition and electron microscopy structure of *Drosophila melanogaster* and human spliceosomal complexes. *Mol. Cell. Biol.* **29**, 281–301 (2009).
  41. Herzel, L. & Neugebauer, K. M. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods* **85**, 36–43 (2015).
  42. Hogg, R., de Almeida, R. A., Ruckshanthi, J. P. D. & O’Keefe, R. T. Remodeling of U2-U6 snRNA helix I during pre-mRNA splicing by Prp16 and the NineTeen Complex protein Cwc2. *Nucleic Acids Research* **42**, 8008–8023 (2014).
  43. Hooks, K. B., Delneri, D. & Griffiths-Jones, S. Intron Evolution in *Saccharomycetaceae*. *Genome Biology and Evolution* **6**, 2543–2556 (2014).
  44. Horowitz, D. S. The mechanism of the second step of pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **3**, 331–350 (2012).
  45. Hossain, M. A. & Johnson, T. L. Using Yeast Genetics to Study Splicing Mechanisms. in *Spliceosomal Pre-mRNA Splicing* (ed. Hertel, K. J.) **1126**, 285–298 (Humana Press, 2014).
  46. Hughes, T. A. Regulation of gene expression by alternative untranslated regions. *Trends in Genetics* **22**, 119–122 (2006).



47. Inada, M. & Pleiss, J. A. Genome-wide approaches to monitor pre-mRNA splicing. *Meth. Enzymol.* **470**, 51–75 (2010).
48. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).
49. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).
50. Katz, Y., Wang, E. T., Airoidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7**, 1009–1015 (2010).
51. Katz, Y. *et al.* Quantitative visualization of alternative exon expression from RNA-seq data. *Bioinformatics* **31**, 2400–2402 (2015).
52. Kelemen, O. *et al.* Function of alternative splicing. *Gene* **514**, 1–30 (2013).
53. Kim, D. H., Edwalds-Gilbert, G., Ren, C. & Lin, R. J. A mutation in a methionine tRNA gene suppresses the prp2-1 Ts mutation and causes a pre-mRNA splicing defect in *Saccharomyces cerevisiae*. *Genetics* **153**, 1105–1115 (1999).
54. Kim, S.-H. & Lin, R.-J. Spliceosome activation by PRP2 ATPase prior to the first transesterification reaction of pre-mRNA splicing. *Molecular and Cellular Biology* **16**, 6810–6819 (1996).
55. Kim, S. W. *et al.* Widespread intra-dependencies in the removal of introns from human transcripts. *Nucleic Acids Research* **45**, 9503–9513 (2017).
56. Kobor, M. S. *et al.* A protein complex containing the conserved Swi2/Snf2-

- related ATPase Swr1p deposits histone variant H2A.Z into euchromatin. *PLoS Biol.* **2**, E131 (2004).
57. Krogan, N. J. *et al.* A Snf2 family ATPase complex required for recruitment of the histone H2A variant Htz1. *Mol. Cell* **12**, 1565–1576 (2003).
  58. Larson, A., Fair, B. J. & Pleiss, J. A. Interconnections Between RNA-Processing Pathways Revealed by a Sequencing-Based Genetic Screen for Pre-mRNA Splicing Mutants in Fission Yeast. *G3 & Genes/Genomes/Genetics* **6**, 1513–1523 (2016).
  59. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15**, R29 (2014).
  60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
  61. Lin, R. J., Newman, A. J., Cheng, S. C. & Abelson, J. Yeast mRNA splicing in vitro. *J. Biol. Chem.* **260**, 14780–14792 (1985).
  62. Liu, H.-L. & Cheng, S.-C. The interaction of Prp2 with a defined region of the intron is required for the first splicing reaction. *Mol. Cell. Biol.* **32**, 5056–5066 (2012).
  63. Long, J. C. & Cáceres, J. F. The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal* **417**, 15–27 (2009).
  64. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, (2014).
  65. Madhani, H. D. & Guthrie, C. Genetic interactions between the yeast RNA

- helicase homolog Prp16 and spliceosomal snRNAs identify candidate ligands for the Prp16 RNA-dependent ATPase. *Genetics* **137**, 677–687 (1994).
66. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
  67. Marinov, G. K. *et al.* From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research* **24**, 496–510 (2014).
  68. Martinez-Campa, C. *et al.* Precise nucleosome positioning and the TATA box dictate requirements for the histone H4 tail and the bromodomain factor Bdf1. *Mol. Cell* **15**, 69–81 (2004).
  69. Maschhoff, K. L. & Padgett, R. A. The stereochemical course of the first step of pre-mRNA splicing. *Nucleic Acids Res.* **21**, 5456–5462 (1993).
  70. Matangkasombut, O. & Buratowski, S. Different sensitivities of bromodomain factors 1 and 2 to histone H4 acetylation. *Mol. Cell* **11**, 353–363 (2003).
  71. Matera, A. G. & Wang, Z. A day in the life of the spliceosome. *Nat. Rev. Mol. Cell Biol.* **15**, 108–121 (2014).
  72. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome research* **9**, 1288–1293 (1999).
  73. Miyoshi, T., Kanoh, J., Saito, M. & Ishikawa, F. Fission yeast Pot1-Tpp1 protects telomeres and regulates telomere length. *Science* **320**, 1341–1344 (2008).
  74. Moll, P., Ante, M., Seitz, A. & Reda, T. QuantSeq 3 [prime] mRNA sequencing for RNA quantification. *Nature Methods* **11**, (2014).
  75. Mordes, D. *et al.* Pre-mRNA splicing and retinitis pigmentosa. *Molecular vision*

- 12**, 1259 (2009).
76. Neugebauer, K. M. On the importance of being co-transcriptional. *Journal of Cell Science* **115**, 3865–3871 (2002).
  77. Nguyen, T. H. D. *et al.* The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature* **523**, 47 (2015).
  78. Nilsen, T. W. & Graveley, B. R. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463 (2010).
  79. Nowrousian, M. Fungal gene expression levels do not display a common mode of distribution. *BMC Res Notes* **6**, 559 (2013).
  80. Oesterreich, F. C. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**, 372–381 (2016).
  81. Ohrt, T. *et al.* Molecular dissection of step 2 catalysis of yeast pre-mRNA splicing investigated in a purified system. *RNA* **19**, 902–915 (2013).
  82. Okoniewski, M. J. & Miller, C. J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC bioinformatics* **7**, 276 (2006).
  83. Oubridge, C., Ito, N., Evans, P. R., Teo, C. H. & Nagai, K. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**, 432–438 (1994).
  84. Pamblanco, M. *et al.* Bromodomain factor 1 (Bdf1) protein interacts with histones. *FEBS Lett.* **496**, 31–35 (2001).
  85. Papasaïkas, P. & Valcárcel, J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. *Trends in Biochemical Sciences* **41**, 33–45 (2016).

86. Pelechano, V., Wei, W., Jakob, P. & Steinmetz, L. M. Genome-wide identification of transcript start and end sites by Transcript Isoform Sequencing, TIF-Seq. *Nat Protoc* **9**, 1740–1759 (2014).
87. Pleiss, J. A., Whitworth, G. B., Bergkessel, M. & Guthrie, C. Transcript Specificity in Yeast Pre-mRNA Splicing Revealed by Mutations in Core Spliceosomal Components. *PLoS Biology* **5**, e90 (2007).
88. Pleiss, J. A., Whitworth, G. B., Bergkessel, M. & Guthrie, C. Rapid, transcript-specific changes in splicing in response to environmental stress. *Mol. Cell* **27**, 928–937 (2007).
89. Podar, M., Perlman, P. S. & Padgett, R. A. The two steps of group II intron self-splicing are mechanistically distinguishable. *Rna* **4**, 890–900 (1998).
90. Price, S. R., Evans, P. R. & Nagai, K. Crystal structure of the spliceosomal U2B''-U2A' protein complex bound to a fragment of U2 small nuclear RNA. *Nature* **394**, 645–650 (1998).
91. Rädle, B. *et al.* Metabolic labeling of newly transcribed RNA for high resolution gene expression profiling of RNA synthesis, processing and decay in cell culture. *J Vis Exp* (2013). doi:10.3791/50195
92. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).
93. Ritchie, D. B., Schellenberg, M. J. & MacMillan, A. M. Spliceosome structure: piece by piece. *Biochim. Biophys. Acta* **1789**, 624–633 (2009).
94. Rockman, M. V. & Kruglyak, L. Genetics of global gene expression. *Nature Reviews Genetics* **7**, 862 (2006).

95. Rossmiller, B., Mao, H. & Lewin, A. S. Gene therapy in animal models of autosomal dominant retinitis pigmentosa. *Mol. Vis.* **18**, 2479–2496 (2012).
96. Rubio-Peña, K. *et al.* Modeling of autosomal-dominant retinitis pigmentosa in *Caenorhabditis elegans* uncovers a nexus between global impaired functioning of certain splicing factors and cell type-specific apoptosis. *RNA* **21**, 2119–2131 (2015).
97. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *Journal of Molecular Biology* **428**, 2623–2635 (2016).
98. Schmidt, C. *et al.* Mass spectrometry–based relative quantification of proteins in precatalytic and catalytically active spliceosomes by metabolic labeling (SILAC), chemical labeling (iTRAQ), and label-free spectral count. *RNA* **20**, 406–420 (2014).
99. Schwarzl, T., Higgins, D. G., Kolch, W. & Duffy, D. J. Measuring Transcription Rate Changes via Time-Course 4-Thiouridine Pulse-Labeling Improves Transcriptional Target Identification. *J. Mol. Biol.* **427**, 3368–3374 (2015).
100. Schwer, B. & Guthrie, C. PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome. *Nature* **349**, 494–499 (1991).
101. Scotti, M. M. & Swanson, M. S. RNA mis-splicing in disease. *Nature Reviews Genetics* **17**, 19–32 (2015).
102. Shalgi, R., Hurt, J. A., Lindquist, S. & Burge, C. B. Widespread Inhibition of Posttranscriptional Splicing Shapes the Cellular Transcriptome following Heat Shock. *Cell Reports* **7**, 1362–1370 (2014).

103. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc Natl Acad Sci U S A* **109**, 1347–1352 (2012).
104. Silverman, E. J. *et al.* Interaction between a G-patch protein and a spliceosomal DEXD/H-box ATPase that is critical for splicing. *Mol. Cell. Biol.* **24**, 10101–10110 (2004).
105. Sims, D., Sudbery, I., Iltott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15**, nrg3642 (2014).
106. Spadaccini, R. *et al.* Biochemical and NMR analyses of an SF3b155-p14-U2AF-RNA interaction network involved in branch point definition during pre-mRNA splicing. *RNA* **12**, 410–425 (2006).
107. Sridharan, V., Heimiller, J. & Singh, R. Genomic mRNA Profiling Reveals Compensatory Mechanisms for the Requirement of the Essential Splicing Factor U2AF. *Molecular and Cellular Biology* **31**, 652–661 (2011).
108. Staley, J. P. & Guthrie, C. Mechanical Devices of the Spliceosome: Motors, Clocks, Springs, and Things. *Cell* **92**, 315–326 (1998).
109. Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A. & Pleiss, J. A. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res.* **43**, 8488–8501 (2015).
110. Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E. & Fairbrother, W. G. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat. Struct. Mol. Biol.* **19**, 719–721 (2012).

111. Tanackovic, G. *et al.* PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Human Molecular Genetics* **20**, 2116–2130 (2011).
112. Tardiff, D. F., Lacadie, S. A. & Rosbash, M. A genome-wide analysis indicates that yeast pre-mRNA splicing is predominantly posttranscriptional. *Mol. Cell* **24**, 917–929 (2006).
113. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
114. Tucker, B. A. *et al.* Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa. *Proceedings of the National Academy of Sciences* **108**, E569–E576 (2011).
115. Umen, J. G. & Guthrie, C. The second catalytic step of pre-mRNA splicing. *RNA* **1**, 869–885 (1995).
116. Uren, P. J., Lee, J. T., Doroudchi, M. M., Smith, A. D. & Horsager, A. A profile of transcriptomic changes in the rd10 mouse model of retinitis pigmentosa. *Molecular vision* **20**, 1612 (2014).
117. Verdel, A. *et al.* RNAi-Mediated Targeting of Heterochromatin by the RITS Complex. *Science* **303**, 672–676 (2004).
118. Vidovic, I., Nottrott, S., Hartmuth, K., Lührmann, R. & Ficner, R. Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell* **6**, 1331–1342 (2000).
119. Vijayraghavan, U., Company, M. & Abelson, J. Isolation and characterization of



- pre-mRNA splicing mutants of *Saccharomyces cerevisiae*. *Genes Dev.* **3**, 1206–1216 (1989).
120. Villa, T. & Guthrie, C. The Isy1p component of the NineTeen complex interacts with the ATPase Prp16p to regulate the fidelity of pre-mRNA splicing. *Genes & Development* **19**, 1894–1904 (2005).
  121. Wan, R. *et al.* The 3.8 Å structure of the U4/U6.U5 tri-snRNP: Insights into spliceosome assembly and catalysis. *Science* **351**, 466–475 (2016).
  122. Wan, R., Yan, C., Bai, R., Huang, G. & Shi, Y. Structure of a yeast catalytic step I spliceosome at 3.4 Å resolution. *Science* **353**, 895–904 (2016).
  123. Wang, Y., Wagner, J. D. & Guthrie, C. The DEAH-box splicing factor Prp16 unwinds RNA duplexes in vitro. *Curr. Biol.* **8**, 441–451 (1998).
  124. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* **10**, 57–63 (2009).
  125. Ward, A. J. & Cooper, T. A. The Pathobiology of Splicing. *J Pathol* **220**, 152–163 (2010).
  126. Wernersson, R., Juncker, A. S. & Nielsen, H. B. Probe selection for DNA microarrays using OligoWiz. *Nature Protocols* **2**, nprot.2007.370 (2007).
  127. Will, C. L. & Luhrmann, R. Spliceosome Structure and Function. *Cold Spring Harbor Perspectives in Biology* **3**, a003707–a003707 (2011).
  128. William Roy, S. & Gilbert, W. The evolution of spliceosomal introns: patterns, puzzles and progress. *Nature Reviews Genetics* **7**, 211–221 (2006).
  129. Wlodaver, A. M. & Staley, J. P. The DExD/H-box ATPase Prp2p destabilizes and proofreads the catalytic RNA core of the spliceosome. *RNA* **20**, 282–294

- (2014).
130. Wolfe, K. H. Origin of the Yeast Whole-Genome Duplication. *PLoS Biol.* **13**, e1002221 (2015).
131. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* aac7629 (2015). doi:10.1126/science.aac7629
132. Yan, C., Wan, R., Bai, R., Huang, G. & Shi, Y. Structure of a yeast activated spliceosome at 3.5 Å resolution. *Science* **353**, 904–911 (2016).
133. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome biology* **5**, R74 (2004).
134. Yung, S. B. & Primm, T. P. Nucleotide Manipulatives to Illustrate the Central Dogma. *J Microbiol Biol Educ* **16**, 274–277 (2015).
135. Zhang, H., Roberts, D. N. & Cairns, B. R. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell* **123**, 219–231 (2005).
136. Zhang, X. *et al.* An Atomic Structure of the Human Spliceosome. *Cell* **169**, 918–929.e14 (2017).
137. Zhang, Z. & Dietrich, F. S. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic Acids Res* **33**, 2838–2851 (2005).
138. Zhuang, Y. & Weiner, A. M. A compensatory base change in U1 snRNA suppresses a 5' splice site mutation. *Cell* **46**, 827–835 (1986).